On Adaptivity and Confounding in Contextual Bandit Experiments

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Multi-armed bandit algorithms minimize experimentation costs required to converge on optimal behavior. They do so by rapidly adapting experimentation effort away from poorly performing actions as feedback is observed. But this desirable feature makes them sensitive to confounding. We highlight, for instance, that popular bandit algorithms cannot address the problem of identifying the best action when day-of-week effects may confound inferences. In response, this paper formulates a general model of contextual bandit experiments with nonstationary contexts, which act as the confounders for inferences and can be also viewed as the distribution shifts in the earlier periods of the experiments. In addition, this general model allows the target distribution or population distribution that is used to determine the best action to be different from the empirical distribution over the contexts observed during the experiments. The paper proposes *deconfounded Thompson sampling*, which makes simple, but critical, modifications to the way Thompson sampling is usually applied. Theoretical guarantees suggest the algorithm strikes a delicate balance between adaptivity and robustness to confounding and distribution shifts. It attains asymptotic lower bounds on the number of samples required to confidently identify the best action — suggesting optimal adaptivity — but also satisfies strong performance guarantees in the presence of day-of-week effects and delayed observations — suggesting unusual robustness.

1 Paper Summary

Multi-armed bandit algorithms are designed to adapt their experimentation rapidly as evidence is gathered. By quickly shifting measurements away from less promising actions or 'arms', they focus measurement effort where it is most useful. This desirable feature can make these same algorithms brittle in the face of delayed observations or confounding factors. We highlight this challenge through an example of a week-long experiment where observations are influenced by specific day-of-week effects.

Example 1 (Day-of-week effects). *In any period $t \in [T] := \{1, \cdots, T\}$, the decision-maker observes context $X_t \in [7]$, selects arm $I_t \in [k]$, and observes a noisy reward R_t that reflects the performance of the chosen arm in the current context. For concreteness, one might imagine that a period corresponds to a customer visiting an online retailer, the context indicates the day of the week, an arm indicates the price set for a particular product, and the reward is the resulting revenue. The context at time t is $X_t = \lceil t/m \rceil$, meaning the first m periods are Sunday, the next m are Monday*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

and so on. Assume that $R_t = \theta_{I_t, X_t} + W_t$ where $W_t \mid \theta, I_t \sim N(0,1)$ is independent Gaussian noise and $\theta \in \mathbb{R}^{7k}$ is an unknown parameter vector that encodes the day and arm specific mean rewards. By intelligently adapting measurement effort, the decision-maker hopes to identify the arm

$$I^*(\theta) = \underset{i \in [k]}{\arg\max} \; \frac{\theta_{i,1} + \cdots + \theta_{i,7}}{7} \tag{1}$$

that maximizes expected revenue if employed throughout the entire week. The goal is to learn a single price and not a sequence of seven prices to charge on separate days of the week. Such predictable price variations might, for instance, lead to unintended strategic customer behavior if implemented across many future weeks.

The decision-maker begins with prior belief under which $\theta \sim N(\mu, \Sigma)$. This might, for instance, arise from a latent variable model where $\theta_{i,x} = \theta^{\text{idio}}_{i,x} + \theta^{\text{arm}}_{i} + \theta^{\text{day}}_{x}$ is determined by an effect $\theta^{\text{idio}}_{i,x}$ that is idiosyncratic to a specific arm and day, an effect θ^{arm}_{i} associated with the chosen arm, and a shared day-of week effect θ^{day}_{x}. Placing an independent normal prior on the idiosyncratic, arm-specific, and day-specific effects induces a structured covariance matrix Σ. When the idiosyncratic terms have large variance, the decision-maker must guard against almost arbitrary non-stationary patterns. If these are believed to have smaller magnitude, the decision-maker may be able rule out some very poor arms early in the experiment.

Day-of-week effects are a standard concern when practitioners run A/B tests [Kohavi et al., 2020], so it is concerning that popular bandit algorithms like Thompson sampling and upper confidence bound [Lattimore and Szepesvári, 2020] fail in this example. The issue is that these algorithms either risk confounding by ignoring contextual information or aim to find the best action for every specific context, which is often not the experimenter's goal (see the discussion of coarse segmentation below). In light of this discussion, it may not be surprising that many real life experiments implement uniformly random arm selection. This experimental design is highly robust, but it is inefficient when there are many arms and some can be quickly identified as inferior. The adaptivity of multi-armed bandit algorithms is important in those settings.

We propose *deconfounded Thompson sampling* (DTS). This method involves simple but critical modifications to Thompson sampling — an algorithm that is widely used in industry in academia. Our results suggest that DTS strikes a delicate balance: it is aggressive in shifting measurement effort away from alternatives that appear inferior while being robust to observed confounders like the day-of-week effects in Example 1.

1.1 A Model of Contextual Bandit Experiments

We formulate a general model of contextual bandit experiments that encompasses Example 1 as a special case. The model captures the defining features of Example 1 including:

- *Coarse segmentation:* The ultimate decision-rule (1) pools together all seven contexts into a single segment over which decisions are held constant. If we think of the customers as belonging to one of seven different groups, then specifying a price for each day would be the most granular segmentation and (1) specifies the coarsest. In settings where the context contains all information available about a customer, coarse segmentation reduces data requirements, reduces the risk of bias, and avoids complex strategic incentives that occur when customers' interactions affect their future service. In practice, products, public policies, and health interventions are often designed to serve a segment of the population (e.g. rural millennials) without being specialized to each individual.
- *Nonstationary confounders:* The experimenter needs to account for day-of-week effects in order to correctly infer which arm is best. They may need to model granular contextual information when performing inference even if they want to employ a coarse segmentation for the decisions they implement. The nonstationary

contexts act as the confounders for inferences and can be also viewed as the distribution shifts in the earlier periods of the experiments.

- *Pure exploration:* In the lingo of the multi-armed bandit literature, what we have described is a "pure-exploration" problem [Bubeck et al., 2009]. In common bandit formulations, experimentation continues indefinitely but is costly only if suboptimal action is selected. In our formulation, one hopes to quickly stop the experimentation process and commit to given strategy for selecting actions going forward. This is natural in settings where the process of experimentation is inherently costly, as it is in clinical trials or many public policy experiments. Even in internet experiments, the dominant workflow involves running a finite length experiment to validate or select among alternatives. After an option is selected, engineering resources might be invested toward productionizing it. One salient feature of the model is that the target distribution or population distribution that is used to determine the best action can be different from the empirical distribution over the contexts observed during the experiments.

We formulate a general model that combines these features. A decision-maker experiments across a sequence of periods. In each, they observe a context vector, select from a finite set of possible actions, and observe a reward whose probability distribution depends on the chosen context and action. After the experimentation process stops, the decision-maker commits to a given strategy for selecting actions going forward. Specifically, they pick among a class of candidate policies, each of which is a rule that prescribes an action for every context. Restricting the class of candidate policies enforces coarse segmentation. The decision-maker's choice is judged by how it performs on average under contexts drawn from a population distribution, effectively capturing how that policy will perform when employed throughout an extremely large number of remaining periods. We assume the population distribution is known, which would be essential if the contexts observed during the experiment are not representative of the distribution anticipated in the future. More generally, web companies typically have rich historical data on their users and should not try to estimate this population's attributes separately in each experiment they run.

1.2 Failure of Popular Bandit Algorithms

The two most popular approaches to (stochastic) multi-armed bandit problems are upper confidence bound (UCB) and Thompson sampling (TS) algorithms [see e.g. Slivkins et al., 2019, Lattimore and Szepesvári, 2020]. UCB selects the arm with the highest UCB on its mean reward. TS is a randomized strategy under which the probability of sampling an arm is "matched" to the posterior probability that arm is optimal. Both algorithm have been applied to a variety of complex and interesting online decision-making problems.

We show that neither TS nor UCB, as usually applied, can address Example 1. In problems with contexts, TS and UCB aim to select an action that could plausibly maximize the expected reward earned in the current context. UCB does this by forming a confidence bound on each arm's performance under the current context and TS performs probability matching with respect to the optimal arm in the current context. These strategies do not gather sufficient information about arms that are suboptimal on the current day but might be optimal throughout the week. They also could waste measurement effort on arms that appear almost certain to offer suboptimal average performance throughout the week. Heuristic versions of TS or UCB that disregard contextual information when performing inference would risk confounding due to un-modeled day-of-week effects.

A potential adaptation of UCB to Example 1 would form UCBs on the weeklong average reward in (1). We show this may sample only a single arm on a given day because UCBs do not diminish until later days are observed. As a result, the data it collects cannot be used to identify the best arm in (1), regardless of the length of the problem's time horizon.

1.3 Deconfounded Thompson Sampling

Our proposed algorithm makes two modifications to Thompson sampling as it is usually defined in contextual bandit problems. The first makes the algorithm suitable for learning about a target policy with coarse segmentation. In the setting of Example 1, rather than perform probability matching with respect to the best action for the current day, it performs probability matching with respect to the arm with best performance throughout the week as in (1). More generally, the proposed algorithm performs probability matching with respect to the action prescribed at the current context by the target policy in the policy class. This idea limits exploration to important distinctions between the candidate policies. The second modification makes the algorithm suitable for pure-exploration problems by adapting the top-two sampling strategy of Russo [2020]. This modification explores suboptimal arms more aggressively by running Thompson sampling until two distinct actions are drawn and then randomly picking among those "top-two". We call this algorithm deconfounded Thompson sampling (DTS). Unlike standard TS, it can control for confounding factors without segmenting its decisions on the basis of those confounders.

1.4 Theoretical Results

It is difficult to give a single theoretical analysis that illuminates all the issues that are relevant in practice. Instead, we focus on a single algorithm and prove three distinct results that stress different capabilities. All results study simple regret [Bubeck et al., 2009], which measures the shortfall in the expected future per-period reward earned by the decision-maker's selected policy relative to the best the best policy in the policy class. We elaborate on the results below:

1. *Robustness to delay and confounding:* Our first result removes the assumption that contexts are drawn i.i.d. For analytical tractability, we assume a Gaussian linear model governs reward observations and focus on a best-arm learning problem, where the goal is to identify the best fixed arm to employ in the future. Example 1 serves as a special case. We study the expected simple regret incurred by DTS, conditioned on an arbitrary sequence of contexts. We provide a bound that depends only on the information contained in the contexts and is completely independent of the order in which they arrive, demonstrating robustness to non-stationary confounders that are modeled by the algorithm. This result also allows for an arbitrary delay in observing reward realizations.

2. *Adapting optimally to the problem instance:* Our next result fixes some arbitrary parameter vector and studies expected simple regret conditioned on this vector being the true draw from nature. This can be thought of as a "frequentist" bound, whereas the previous two were "Bayesian." This section again imposes the assumption that contexts are drawn i.i.d. and, for analytical tractability, again assumes a Gaussian linear model governs reward observations and focuses on a best-arm learning problem. A fundamental lower bound shows how the expected sample size of an adaptive experiment must grow in order to guarantee some vanishing level of simple regret. The sampling requirements are milder for problem instances where some arms are far from optimal and can be effectively discarded with few samples. We prove that DTS meets attains this asymptotic lower bound. In this sense it optimally adapts its experimentation to the problem instance.

It may not be difficult to design an algorithm that attains one of the results above. It is remarkable, however, that these distinct properties are satisfied simultaneously by one simple heuristic algorithm. Attaining both simultaneously seems to require a delicate balance between robustness and adaptivity.

References

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

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

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Chao Qin, Diego Klabjan, and Daniel Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 2017:5382–5392, 2017.

Daniel Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 2020.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

A Problem Formulation

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

We assume the decision-maker has access to a probability distribution w over contexts that encodes the frequency with which they expect contexts to occur in the future. We call this either the *target distribution* or the *population distribution*, where the latter suggests that w denotes the characteristics of a population of individuals. If employed across a large number of future periods, arm i would generate average reward

$$\mu(\theta, i, w) := \sum_{x \in \mathcal{X}} w(x) \langle \theta^{(i)}, x \rangle = \langle \theta^{(i)}, X_{\text{pop}} \rangle \quad \text{where} \quad X_{\text{pop}} := \sum_{x \in \mathcal{X}} w(x) x. \tag{2}$$

In Example 1, X_{pop} is the vector $(1/7, \cdots, 1/7)$ and $\mu(\theta, i, w)$ is the average that appears in Equation (1). If the decision-maker knew θ, the optimal arm to employ in the future would be $I^* = I^*(\theta) \in \arg\max_{i \in [k]} \mu(\theta, i, w)$.

For technical or notational convenience, we make several additional assumptions. First, we assume $\mathcal{X}$ is finite (though possibly enormous), which allows us later to analyze a lower bound on performance that is expressed through a finite dimensional optimization problem. Second, we assume that the arm-specific parameters $\theta^{(i)}$ are drawn independently across arms, allowing us to track beliefs separately across arms in the analysis. Assume also that the prior covariance matrix $\Sigma_{1,i}$ is the same for each arm i and is positive definite. We denote this by Σ_1.

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

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

$$\Sigma_{t,i} = \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}\{I_\ell = i\} X_\ell X_\ell^\top \right)^{-1} \qquad \mu_{t,i} = \Sigma_{t,i} \left(\Sigma_1^{-1} \mu_{1,i} + \sum_{\ell=1}^{t-L} \mathbb{1}\{I_\ell = i\} X_\ell R_\ell \right). \tag{3}$$

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

$$s_{t,i}^2 = X_{\text{pop}}^\top \Sigma_{t,i} X_{\text{pop}} \qquad m_{t,i} = \langle X_{\text{pop}}, \mu_{t,i} \rangle. \tag{4}$$

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

Performance measures. Let $H_T^+ = (X_{1:T}, I_{1:T}, R_{1:T})$ denote all information generated by a T–period experiment, including the delayed reward outcomes. The non-negative random variable

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

measures the shortfall in future performance caused by selecting the greedy decision at time T by $\hat{I}_T^+ \in \arg\max_{i\in[k]} \mathbb{E}\left[\mu(\theta, i, w) \mid H_T^+\right]$ with only the incomplete information about θ accrued after T measurements. We call Δ_T the *simple regret* at time T, after Bubeck et al. [2009]. Having in mind policy decisions where $\mu(\theta, i, x)$ denotes the utility generated for an individual with features x, Athey and Wager [2021] call this the *utilitarian regret*. Notice that the decision $\hat{I}_T^+$ can be made using the full results of the experiment H_T^+ while a measurement decision I_t must be made in real-time based on partial information H_t.

The goal in the problem, informally, is to experiment intelligently so that simple regret is small after using as few measurements as possible. This objective is can be formalized in several ways. We focus on two ways of studying performance that allow for clear analytical insight into specific properties of deconfounded Thompson sampling:

1. (Fixed budget and Bayesian) In Section D, we study the expected simple regret $\mathbb{E}\left[\Delta_T \mid X_{1:T} = x_{1:T}\right]$ at some finite time T, conditioned on the sequence of realized contexts $X_{1:T} := (X_1, \cdots, X_T)$ taking on some specific value. This expectation integrates over most randomness in the problem, including over the prior distribution, and emphasizes dependence on the observed contexts and their order. Our goal in this section is to show that deconfounded TS satisfies an important robustness property other adaptive algorithms do not: roughly speaking, we have a result of the form $\mathbb{E}\left[\Delta_T \mid X_{1:T} = x_{1:T}\right] \leqslant \tilde{O}(\sqrt{k/T})$, where the big-$O$ hides a natural dependence on the second moment $\frac{1}{T}\sum_{t=1}^{T} x_t x_t^\top$ but has no dependence on context order.
2. (Adaptive stopping and frequentist) In Section E, we aim to verify that the algorithm adapts its measurement effort optimally, in an appropriate sense, as it learns about the true problem instance. To do this, we study performance conditional on the unknown parameter θ but integrate over the distribution of the contexts $(X_1, X_2, \cdots)$, which we assume to drawn i.i.d. Following the style of result in Russo [2020], Glynn and Juneja [2004] or Kaufmann et al. [2016], we would hope to show that $\mathbb{E}\left[\Delta_T \mid \theta\right]$ goes to zero at an exponential rate T grows, and that the problem-dependent exponent is in appropriate sense the best-possible among adaptive algorithms. This is called the "fixed-budget" formulation in the literature on the best-arm identification literature, because there is a hard constraint on the number of samples (i.e T) that can be collected that must be satisfied with probability one. Unfortunately, the sharp asymptotic limits in that setting are poorly understood, even in problems without contexts. We instead look at formulations in which there is a soft-constraint on the number of measurements. There we study performance at a adaptively chosen stopping time τ, which essentially, stops once the posterior expectation of simple regret is small. We study the combined cost $\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta\right]$ as $c \to 0$, measuring the *expected* number of samples required to deliver vanishing simple regret. This formulation follows classic work of Chernoff et al. [1959]; very similar results follow if one instead imposes a constraint on the simple regret or the probability of incorrect selection, which is called the "fixed-confidence" setting. Kaufmann et al. [2016].

B Deconfounded Thompson Sampling

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

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

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

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

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

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

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

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

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

Efficient computation. Following conventional implementation of Thompson sampling, a generic approach sampling from α_t, is to sample a parameter vector $\tilde{\theta}$ from the posterior distribution of θ and then to find the arm $\arg\max_{i\in[k]} \mu(\tilde{\theta}, i, w)$ that is best under this sample.The structure of Gaussian linear belief models allows for an even cleaner implementation of DTS. Because the population average reward of arm i, $\mu(\theta, i, w)$, has a Gaussian posterior with posterior parameters given in (4), one can directly perform inference on the population average rewards.

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

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

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

Input prior parameters $(\mu_{1,i}, \Sigma_{1,i})_{i\in[k]}$, population weights X_{pop} and noise variance σ^2.
for $t = 1, 2, \cdots$ **do**
  Sample $\nu_i \sim N(m_{t,i}, s^2_{t,i})$ for $i \in [k]$ and set $I_t^{(1)} = \arg\max_{i\in[k]} \nu_i$;
  do
    Sample $\nu_i \sim N(m_{t,i}, s^2_{t,i})$ for $i \in [k]$ and set $I_t^{(2)} = \arg\max_{i\in[k]} \nu_i$;
  while $I_t^{(1)} = I_t^{(2)}$;
  Flip coin $C_t \in \{0, 1\}$ with bias $\mathbb{P}(C_t = 1) = \beta_t$;
  Play arm $I_t = I_t^{(1)} C_t + I_t^{(2)}(1 - C_t)$;
  Gather delayed observation $o = (I_{t-L}, X_{t-L}, R_{t-L})$.;
  Calculate posterior parameters $m_{t+1,i}$, $s^2_{t+1,i}$ for $i \in [k]$ according to (4) to reflect o;
  Calculate new tuning parameter β_{t+1} if using adaptive tuning;
end

C Failure of Alternative Bandit Algorithms

This section provides examples showing that alternative bandit algorithms can fail for simple examples within the scope of our problem formulation. Most interesting, perhaps, is that a deconfounded UCB algorithm cannot address a simplified version of the example with day-of-week effects described in the introduction. Past theory on TS highlights connections to UCB, so any theory of DTS in that example will need to push well beyond current understanding. We also show the failure of a context unaware algorithm and a usual contextual bandit algorithm.

What does it mean that these algorithms 'fail'? We show formally that simple regret is bounded as $\mathbb{E}[\Delta_T] \geqslant c$, where c is some absolute numerical constant that does not depend on T. Regardless of the time dedicated to experimentation, the data the collected by these algorithms is inadequate and cannot be used to make near-optimal decisions. The examples we describe are meant to give insight into what can go wrong with alternative algorithms and the subtleties of designing an algorithm like DTS. They are purposefully simplistic.

C.1 Deconfounded UCB

Consider the following simplification of Example 1. Here, there are two contexts instead of seven and we restrict to the case of two actions.

Example 2 (Day of week effects)**.** *The context set is $\mathcal{X} = \{1, 2\}$ and there are $k = 2$ arms. The reward at time t is $R_t = \theta_{X_t}^{(I_t)} + W_t$ where each $\theta_x^{(i)}$ is independent and Gaussian and $W_t \sim N(0, \sigma^2)$ is i.i.d Gaussian noise. Observations are not subject to delay (i.e $L = 1$). The the goal is to identify the best arm under equal context weights w:*

$$I^* = \arg\max_{i \in [2]} \frac{\theta_1^{(i)} + \theta_2^{(i)}}{2}.$$

The context sequence is non-random, with $X_t = 1$ for $t \leqslant \lfloor T/2 \rfloor$, $X_t = 2$ for $t > \lfloor T/2 \rfloor$.

Consider a UCB analogue of our Thompson sampling based algorithm. Reflecting that the true goal is to select an arm with strong performance throughout the week, not on a specific day, it plays the arm with the highest UCB on its average performance throughout the week:

$$I_t \in \arg\max_{i \in [k]} m_{t,i} + z \cdot s_{t,i} \quad \text{for all } t \in \mathbb{N}. \tag{5}$$

where $m_{t,i}$ and $s_{t,i}$ are defined in (4) and $z > 0$ is a tuning parameter. When $z = 1.645$, the term $m_{t,i} + z \cdot s_{t,i}$ measures the 95% quantile of the posterior distribution. Like DTS, this can be thought of as a *deconfounded* UCB, which still selects the arm with the highest upside but accounts for observed confounders when performing inference.

The next result shows formally that deconfounded UCB fails to collect adequate data, regardless of the length of the time horizon. The issue is that the UCB in (5) is sometimes higher for action 2 for each of the first $T/2$ periods. Action 1 is then never sampled in context 1, so learning is incomplete. This holds true regardless of how z is set and holds for time dependent tuning parameters. The issue is that, unlike common bandit settings, UCBs do not diminish when actions are repeatedly sampled in a single context.

Lemma 1. *Consider Example 2. Suppose that the components of the vector $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$ are independent with $\theta_x^{(1)} \sim N(0, 1)$ and $\theta_x^{(2)} \sim N(0, 2)$ for $x \in \{1, 2\}$, and $\sigma^2 = 0$. If* (5) *holds, then there is an absolute numerical constant $c > 0$ such that $\mathbb{E}[\Delta_T] \geqslant c$ for any $T \in \mathbb{N}$.*

C.2 Context Unaware Algorithms

Our next example highlights the risk of confounding for an algorithm that does not model day-of-week effects when performing inference. We set $\tilde{s}_{t,i}^2 = \left(1 + \sigma^{-2} \sum_{\ell=1}^{t-1} \mathbb{1}(I_\ell = i)\right)^{-1}$ and $\tilde{m}_{t,i} = \tilde{s}_{t,i}^2 \left(\sum_{\ell=1}^{t-1} \mathbb{1}(I_\ell = i) R_\ell\right)$. We define these expressions for $\sigma^2 = 0$ by taking the limit as $\sigma^2 \downarrow 0$. In particular, we set $\tilde{s}_{t,i}^2 = 0$ if arm i has been played previously and $\tilde{m}_{t,i}$ to be 0 if arm i was never played previously and to be the the empirical average reward otherwise. These are the posterior updating equations if $\theta_1^{(i)} \sim N(0, 1)$ and the algorithm (incorrectly) ignores day of week effects and assumes $\theta_2^{(i)} = \theta_1^{(i)}$ almost surely.

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

$$I_t = \arg\max_{i \in [2]} \nu_{t,i} \quad \text{where} \quad \nu_{t,i} \mid H_t \sim N(\tilde{m}_{t,i}\,,\, \tilde{s}^2_{t,i}). \tag{6}$$

In the above equation $\nu_{t,1}$ and $\nu_{t,2}$ are sampled independently. The next lemma formalizes that this algorithm risks confounding. The same result applies to a context unaware form UCB, which forms UCBs based on $\tilde{m}_{t,i}$ and $\tilde{s}^2_{t,i}$. A context-unaware top-two TS algorithm fails in a similar way in problems with more than two actions.

Lemma 2 (Failure of context unaware Thompson sampling)**.** *Consider Example 2. 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 $x \in \{1,2\}$, and $\sigma^2 = 0$. If* (6) *holds then there exists an absolute numerical constant $c > 0$ such that for all $T \in \mathbb{N}$, $\mathbb{E}[\Delta_T] \geqslant c$.*

C.3 Contextual Bandit Algorithms

The goal in our formulation is to select among a very restricted set of decision-rules: those that choose a common action, irrespective of context. Experimentation should be tailored to this objective. Here, we give insight into potential failures when the 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.

Example 3 (A mainstream action)**.** *Consider a problem with $k = 3$ arms and 2 contexts given as $\mathcal{X} = \{1,2\}$. The population distribution w is uniform over $\mathcal{X}$ and $(X_t)_{t \in \mathbb{N}}$ are drawn i.i.d from w. The components of the parameter vector $\theta = (\theta_0, \theta_1, \theta_2)$ are drawn independently with $\theta_0 \sim \text{Uniform}([0,1])$ and $\theta_x \sim \text{Uniform}(\{1,2\})$ for $x \in [2]$. Rewards are noiseless, with $R_t = \mu(\theta, I_t, X_t)$. Observations are not subject to delay (i.e. $L = 1$). Action 3's performance is insensitive to the context, and it always generates mean-reward $\mu(\theta, 3, x) = \theta_0$. Actions 1 and 2 generate mean rewards in context $x \in \mathcal{X}$ given by*

$$\mu(\theta, 1, x) = 1/2 + (1/2)\mathbb{1}(\theta_x = 1), \quad \mu(\theta, 2, x) = 1/2 + (1/2)\mathbb{1}(\theta_x = 2).$$

The next lemma formalizes that contextual Thompson sampling, which selects an action according to the posterior probability it is the optimal action for the current context, has simple regret that does not vanish even as the horizon grows. The same result applies to appropriate contextual versions of UCB. The simple reason is that action 3 is never sampled, because it does not maximize the reward in either context. This means no information about θ_0 is gathered and the decision-maker cannot determine whether action 3 is the best arm to select. If the goal is to identify the best policy within a restricted class, the exploration algorithm needs to be designed so that it gathers the right information for this task. The proof follows from this argument and is omitted for brevity.

Lemma 3. *(Failure of contextual Thompson sampling) Suppose that $\mathbb{P}(I_t = i \mid H_t) = \mathbb{P}(I^*(\theta; X_t) = i \mid H_t)$ for each $i \in [k]$. Under Example 3, there is an absolute numerical constant $c > 0$ such that for all $T \in \mathbb{N}$, $\mathbb{E}[\Delta_T] \geqslant c$.*

D Result 1: Robustness to Delay and Confounding

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

D.1 Performance Guarantee

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

$$V(X_{1:T}) = X_{\text{pop}}^{\top} \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^{T} X_t X_t^{\top} \right)^{-1} X_{\text{pop}}. \tag{7}$$

The matrix $\left(\Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^{T} X_t X_t^{\top} \right)^{-1}$ appearing in (7) would be the posterior covariance matrix of $\theta^{(i)}$ at the end of the experimentation horizon if that arm were played in every period. We similarly think of $V(X_{1:T})$ as the posterior variance $\mathrm{Var}(\mu(\theta, i, w) \mid H_T^{+})$ of the population effect of arm i if we observed the reward it generated in every period of the experiment. Notice that what makes the day-of-week effects in Example 2 challenging is *the order* in which contexts arrive. But observing a single arm throughout the entire experiment would be informative, and so $V(X_{1:T})$ would be small if T were large.

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

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

Proposition 1. *Suppose that $\|X_t\|_2 \leqslant 1$ almost surely for $t \in \mathbb{N}$. If DTS applied with tuning parameters satisfying $\inf_{t \in \mathbb{N}} \beta_t \geqslant 1/2$ almost surely, then for any $T \in \mathbb{N}$,*

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

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

Under a natural condition that ensures the context sequence contains sufficient information about the population distribution, the next corollary of Proposition 1 gives a simple-regret bound that scales as $\tilde{O}\left(\sqrt{k/T}\right)$. Notice that this result is nearly-independent of the dimension of the linear model d. If $X_t \sim w$, then $\mathbb{E}\left[X_t X_t^{\top}\right] = X_{\text{pop}} X_{\text{pop}}^{\top} + \mathrm{Cov}(X_t)$. In this sense, if context vectors have high variance in every direction, the bound $\frac{1}{T} \sum_{t=1}^{T} x_t x_t^{\top} \succeq X_{\text{pop}} X_{\text{pop}}^{\top}$ may underestimate the information they provide and make this corollary conservative.

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

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

where ι is given in Proposition 1.

E Result 2: Adaptivity and Asymptotic Optimality

Like most popular multi-armed bandit algorithms, DTS allocates measurement effort adaptively. As time proceeds, it learns about the quality of different policies or arms. By shifting most measurements away from clearly inferior alternatives, it focuses experimentation effort where it is most useful. The previous section showed, although adaptivity makes other natural algorithms brittle in the face of nonstationary confounders, DTS has certain robustness guarantees. This section aims to formalize that DTS also adapts its measurement effort very effectively and, in a sense, *optimally* in a meaningful special case of our formulation.

E.1 Asymptotic Optimality Notion

We assess how effectively the algorithm uses its limited measurements, essentially, by understanding the rate at which simple regret decays as measurements are gathered. Among the several natural ways of studying this, we focus on one that allows for a sharp and enlightening asymptotic theory. We build on asymptotic limits of sequentially designed experiments that have been understood since classic work of Chernoff et al. [1959]. We allow the decision-maker to decide adaptively when to stop collecting measurements. The total cost incurred is $c\tau + \Delta_\tau$, where τ denotes the chosen stopping time, $c > 0$ is a cost per-period of experimentation, and Δ_τ is the simple-regret of the final decision.

We study the expected cost incurred under problem instance θ_0, given by

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right]. \tag{8}$$

In this section, we focus on the parameter class

$$\Theta \triangleq \left\{\theta \in \mathbb{R}^{dk} \;:\; \underset{i\in[k]}{\arg\max}\, \mu(\theta, i, w) \text{ is unique}\right\}.$$

In other words, each parameter in Θ corresponds to a problem instance with a unique best arm under the population distribution.

Sharp results can be established through asymptotic analysis as c tends to zero. This is a regime where the cost of gathering one more observation is negligible relative to the cost committing to a sub-optimal final decision. It arises naturally if one imagines the final decision will later be implemented for a very large number of periods. We establish a kind of uniform optimal guarantee, roughly showing that DTS minimizes (8) to first-order asymptotically *for every specific instance* θ_0. This is only possible under an algorithm that tailors its experimentation optimally to θ_0 as information is gathered.

This theory requires an appropriate stopping rule is used. Attaining the exact optimal constant also requires tuning the β_t parameter as information about θ_0 is acquired. We discuss how this can be done with low computational cost and also discuss robustness with some non-adaptive choices of β.

Our result directly builds on previous analyses that have established similar results for top-two sampling rules [Russo, 2020, Qin et al., 2017]. It may be surprising, however, that these results extend to a contextual setting with linear models, given that more complex exploration rules are often required to attain asymptotic optimality results in problems with parametric dependencies [See e.g. Lattimore and Szepesvari, 2017].

E.2 Notation

Recall the definitions of $m_{t,1} = \left\langle X_{\text{pop}}, \mathbb{E}\left[\theta^{(i)} \mid H_t\right]\right\rangle$ and $s_{t,i}^2 = \text{Var}\left(\mu(\theta, i, w) \mid H_t\right)$ given in . Since the analysis in this section is conditioned on θ, it is helpful to develop analogous notation for these quantities when an improper prior is used. Were an improper prior is used, $\mu_{t,i}$ and $\sigma_{t,i}^2$ would have the formulas:

$$\hat{m}_{t,i} = X_{\text{pop}}^{\top}\left[\sum_{\ell=1}^{t} \mathbb{1}(I_\ell = i) X_\ell X_\ell^{\top}\right]^{-1} \sum_{\ell=1}^{t} \mathbb{1}(I_\ell = i) X_\ell R_t$$

$$\hat{s}_{t,i}^2 = X_{\text{pop}}^{\top}\left[\sigma^{-2}\sum_{\ell=1}^{t} \mathbb{1}(I_\ell = i) X_\ell X_\ell^{\top}\right]^{-1} X_{\text{pop}}.$$

Note that $\hat{m}_{t,i}$ is simply the inner product of X_{pop} with the least-squares estimate for $\theta^{(i)}$. If the chosen arms $\{I_\ell\}$ were fixed in advance rather than selected adaptively, then $\hat{s}_{t,i}^2$ would be the formula for the sampling variance of $\hat{m}_{t,i}$.

It will be important to measure the strength of evidence that one arm outperforms another in the population. For this purpose, consider the natural test of the null hypothesis $\mu(\theta, i, w) \neq \mu(\theta, j, w)$. The classic test would be based on the z-score for the difference in means,

$$Z_{t,i,j} := \frac{m_{t,i} - \mu_{t,j}}{\sqrt{s_{t,i}^2 + s_{t,j}^2}}. \tag{9}$$

Each $Z_{t,i,j}$ follows a normal distribution with unit variance when $I_1, \cdots, I_{t-1}$ are chosen non-adaptively.

E.3 Lower Bound

Let $\mathcal{S} = \{v \in \mathbb{R}_+^k : \sum_{i=1}^k v_i = 1\}$ denote the $k-1$ dimensional probability simplex. Define the complexity measure $\Gamma(\theta)$ by

$$\Gamma(\theta)^{-1} = \sup_{p:\mathcal{X}\to\mathcal{S}} \min_{i\neq I^*} \frac{1}{2\sigma^2} \frac{\left(\mu(\theta, I^*, w) - \mu(\theta, i, w)\right)^2}{X_{\text{pop}}^{\top}\left(\mathbb{E}\left[p(X_1, I^*)X_1X_1^{\top}\right]^{-1} + \mathbb{E}\left[p(X_1, i)X_1X_1^{\top}\right]^{-1}\right) X_{\text{pop}}} \tag{10}$$

where p is a stochastic kernel, which is associates any $x \in \mathcal{X}$ with an element $p(x, \cdot) \in \mathcal{X}$. In this optimization problem, we imagine the experimenter the action at time t by sampling from $p(\cdot|X_t)$. The problem (10) seeks a measurement rule p that maximizes the growth rate of the *minimal* z-score $\min_{j\neq I^*} Z_{t,I^*,j}$. One can then think of $\Gamma(\theta)^{-1}$ as determining a fundamental limit on the rate at which an experimenter can gather evidence against all alternative arms. A peculiar feature of this complexity term is that actually optimizing over p as (10) prescribes would requiring knowing θ, which is circular as uncertainty about θ is point of experimenting in the first place. Nevertheless this complexity measure serves to produce a valid lower bound, as evidenced by the next proposition. The lower bound here applies ideas that has been known since Chernoff et al. [1959], but our proof specifically applies inequalities of Kaufmann et al. [2016].

Proposition 2. *If*

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] \leqslant O(c\log(1/c)) \quad \textit{for all } \theta_0 \in \Theta,$$

as $c \to 0$, then

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] \geqslant \Gamma(\theta_0)[c + o(1)]\log(1/c) \qquad \textit{for all } \theta_0 \in \Theta. \tag{11}$$

The idea of this lower bound is that any algorithm that outperforms (11) on some instance must attain a loss an *an order of magnitude* larger on some other instance. The result is shown, essentially, by establishing that any algorithm with uniformly vanishing simple regret — meaning $\mathbb{E}[\Delta_\tau \mid \theta = \theta_0] = o(1)$ for all θ_0 — must gather an expected number of samples that scales as $\mathbb{E}[\tau \mid \theta = \theta_0] \geqslant \Gamma(\theta_0)(\log(1/c) + o(1))$.

E.4 Optimality of Context Independent Sampling frequencies

The lower bound above turns out to be tight. It is matched by adaptive algorithms that learn as information is gathered to adjust their measurement proportions rapidly enough toward proportions that attain the maximum in (10). As such, the form of the solution is of particular importance. Here we show a striking simplification. The maximal information rate in (10) can be attained by context independent allocation, which samples each arm with a probability that is independent of context. One we reduce a context-independent allocations, it is easy to characterize the solution in terms of the first-order necessary conditions of optimality. Equations (12) and (13) are known for problems without contexts [Glynn and Juneja, 2004].

Lemma 4 (Optimality of context independent sampling frequencies)**.** *Suppose $\mathcal{X}$ is finite. There exists a vector $p^* = p^*(\theta) \in \mathcal{S}$ such that the rule given by $p(x, i) = p^*_i$ for all $x \in \mathcal{X}$ attains the supremum in* (10)*. The vector p^* is the unique solution to the k nonlinear equations:*

$$\frac{\mu(\theta, I^*, w) - \mu(\theta, i, w)}{\sqrt{(p^*_{I^*})^{-1} + (p^*_i)^{-1}}} = \frac{\mu(\theta, I^*, w) - \mu(\theta, j, w)}{\sqrt{(p^*_{I^*})^{-1} + (p^*_j)^{-1}}} \qquad \forall i, j \neq I^* \tag{12}$$

$$p^*_{I^*} = \sqrt{\sum_{i \neq I^*} (p^*_i)^2} \tag{13}$$

Then Equation (10) *becomes*

$$\Gamma(\theta)^{-1} = \frac{1}{2\|X_{\text{pop}}\|_A} \frac{(\mu(\theta, I^*, w) - \mu(\theta, i, w))^2}{(p^*_{I^*})^{-1} + (p^*_i)^{-1}} \quad \forall i \neq I^*$$

where $A = \sigma^2 \left(\mathbb{E}[X_1 X_1^\top]\right)^{-1}$.

We refer to equation (12) as imposing *information balance*. It essentially ensures that the z-scores $Z_{t,I^*,j}$ grow at an equal rate for arms $j \neq I^*$, balancing the evidence against each suboptimal arm.

E.5 Adaptive Tuning

We will show that DTS automatically gathers information in a manner that satisfies an information balance property like Equation (12). By shifting measurement effort away from clearly inferior arms and toward those that could more plausibly be the best arm, the algorithm automatically balances the rate of information acquisition. The precise fraction of measurement effort that (13) suggests should be assigned to the optimal arm is not satisfied automatically, however. In order to do that, the tuning parameter β_t needs to be adjusted properly.

Algorithm 2: Adaptive Tuning Algorithm

Input posterior means of expected reward $(m_{t,i})_{i\in[k]}$.
if $\hat{I}_t$ *is not unique* **then**
  Set $\beta_t = \beta_{t-1}$
end
else
  Obtain the unique optimal solution $x \in \mathcal{S}$ of the empirical version of Equations (12) and (13) with $(\mu(\theta, i, w))_{i\in[k]}$ and I^* replaced by $(m_{t,i})_{i\in[k]}$ and $\hat{I}_t$, respectively:

$$\frac{m_{t,\hat{I}_t} - m_{t,i}}{\sqrt{x_{\hat{I}_t}^{-1} + x_i^{-1}}} = \frac{m_{t,\hat{I}_t} - m_{t,j}}{\sqrt{x_{\hat{I}_t}^{-1} + x_j^{-1}}} \qquad \forall i, j \neq \hat{I}_t \tag{14}$$

$$x_{\hat{I}_t} = \sqrt{\sum_{i\neq\hat{I}_t} x_i^2} \tag{15}$$

  Set $\beta_t = x_{\hat{I}_t}$
end

Efficient Implementation of the Tuning Algorithm For each $i \in [k]$, we define $\Delta_{t,i} \triangleq m_{t,\hat{I}_t} - m_{t,i}$. Equation (14) implies there exists y such that

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

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

$$\frac{x_{\hat{I}_t}}{x_i} = \Delta_{t,i}^2 y - 1, \quad \forall i \neq \hat{I}_t. \tag{16}$$

Together with Equation (15), Equation (16) implies

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

We can solve this fixed-point equation for y using, for example, bisection search or Newton's method. Notice that if Newton's method is used, one may wish to save the value of y solved in the previous time period, which provides an effective initial point for finding an updated value of y. Finally $\sum_{i\in[k]} x_i = 1$ and Equation (16) imply

$$x_{\hat{I}_t} = \frac{1}{1 + \sum_{i\neq\hat{I}_t} \left(\Delta_{t,i}^2 y - 1\right)^{-1}},$$

which is the value assigned to β_t.

E.6 DTS Attains the Lower Bound

We now show that when β_n is tuned as suggested in the previous section, and an appropriate stopping rule is employed, DTS matches the fundamental lower bound in (11).

We consider the empirical selection rule

$$\hat{I}_t \in \arg\max_i \hat{\mu}_{t,i} \tag{17}$$

that selects the arm with highest performance under a least-squares estimate. Similar results can be developed if the Bayes selection rule were used instead, which essentially uses ridge-regression rather than least-squares.

Developing stopping rules is itself an area of active research. We do not try to advance that literature, and instead focus on a very simple candidate that is sufficient for the results we wish to prove. Recall that the z-score $Z_{t,\hat{I}_t,j}$ measures the strength of evidence that arm $\hat{I}_t$ outperforms arm j in the population. The stopping rule

$$\tau = \inf\left\{ t \in \mathbb{N} : \min_{j \neq \hat{I}_t} Z_{t,\hat{I}_t,j} \geqslant \gamma_t \right\} \quad \text{where} \quad \gamma_t = \Phi^{-1}\left(1 - \frac{c}{t^2 k}\right), \tag{18}$$

stops at the first time all z-scores exceed a threshold. The threshold was picked to ensure a probability of incorrect selection less than c. The specific choice of γ_t is based on a Bonferroni correction to account for multiple hypothesis testing and could likely be reduced through more granular analysis.

The next proposition gives two upper bounds.

Proposition 3. *Under the selection rule* (17)*, the stopping rule* (18)*, and allocation rule DTS with* β_t *defined by Algorithm 2, for any* $\theta_0 \in \Theta$*,*

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] \leqslant \Gamma(\theta_0)[c + o(1)] \log(1/c) \qquad \textit{as } c \to 0.$$

If instead the allocation rule is DTS with fixed $\beta = 1/2$*, then for any* $\theta_0 \in \Theta$*,*

$$\mathbb{E}\left[c\tau + \Delta_\tau \mid \theta = \theta_0\right] \leqslant 2\Gamma(\theta_0)[c + o(1)] \log(1/c) \qquad \textit{as } c \to 0.$$

This shows that DTS with adaptively tuned $\{\beta_t\}$ attains the exact optimal constant defined in Equation (10), which matches the lower bound in (11). In addition, DTS with non-adaptive choice of $\beta = 1/2$ achieves near-optimal statistical guarantee while reduces computational cost.