On Adaptivity and Confounding in Contextual Bandit Experiments

Chao Qin CQIN22@GSB.COLUMBIA.EDU
Columbia University

Daniel Russo DJR2174@GSB.COLUMBIA.EDU
Columbia University

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 factors or delayed feedback. 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. 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.

Keywords: contextual bandit, best-arm identification, deconfounded Thompson sampling, randomized controlled trials

1. Extended Abstract

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 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.

There are at least two reasons this may be preferable to the goal of identifying the seven arms that maximize expected revenue on each specific day. First, it requires less data and therefore limits experimentation costs. Second, it avoids undeseriable strategic behavior from customers that might result if prices vary predictably throughout the week.

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. In Example 1, the population distribution is simply uniform, as reflected in (1). 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. This model requires grappling with a number of realistic challenges, including delayed feedback, leveraging prior knowledge, and distributions shift.

Inferences are enabled through a statistical model. A likelihood function specifies the distribution of rewards as a function of the chosen arm, the context, and an unknown and possibly high dimensional parameter vector. Uncertainty about these parameters induces uncertainty about which policy to select at the end of the experiment. We model this uncertain parameter as a random variable drawn from some prior distribution. This has two distinct advantages. First, as discussed further below, it gives practitioners a natural way of leveraging data from outside the experiment. Second, since the posterior distribution gives an unabigious way of performing inference, this choice allows us to focus our research on the principles of sequential decision making and not e.g. the design of confidence intervals.

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). In classical multi-armed bandit problems, UCB algorithms can be viewed as an asymptotic approximation to an exact dynamic programming solution of Gittins (1979). See Chang and Lai (1987); Gutin and Farias (2016); Russo (2021) for further discussion. 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. It was proposed in a simplified form almost ninety years ago (Thompson, 1933), but only in the past decade has it become a leading approach to exploration in industry and academia. 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.

While DTS requires only simple algorithmic modifications to TS, substantial new thought was required to understand that these modifications were sufficient. Previous work on TS has emphasized connections to UCB algorithms (Agrawal and Goyal, 2013; Russo and Van Roy, 2014; Abeille and Lazaric, 2017). But, as described above, the natural form of Deconfounded UCB fails completely in Example 1. One of this paper's primary insights is that the randomized nature of DTS allows it to succeed where UCB does not.

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. The analysis in this section seems to offer substantial innovation. Our proofs use inverse propensity weights implicitly in the analysis, highlighting the importance of randomization and offering an interesting connection to the causal inference literature.

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. Our results in this section are in the style of the classic work by Chernoff (1959) on the asymptotics of sequential experimental design and its application in best-arm identification by Russo (2020); Garivier and Kaufmann (2016). 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

Marc Abeille and Alessandro Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.

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

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.

Fu Chang and Tze Leung Lai. Optimal stopping and dynamic allocation. *Advances in Applied Probability*, pages 829–853, 1987.

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

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Eli Gutin and Vivek F. Farias. Optimistic gittins indices. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3161–3169, 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 Szepesvári. *Bandit Algorithms.* Cambridge University Press, 2020.

Daniel Russo. Simple Bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.

Daniel Russo. A note on the equivalence of upper confidence bounds and gittins indices for patient agents. *Operations Research*, 69(1):273–278, 2021.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

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

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

Appendix A.

Here is the link to the full paper:
https://anonymous.4open.science/r/ReALML2022-CB54/ReALML2022_full_paper.pdf.