Simple Regret Minimization for Contextual Bandits

Aniket Anand Deshmukh [*1] **Srinagesh Sharma** [*2] **James W. Cutler** [3] **Mark B. Moldwin** [4] **Clayton Scott** [5]

Abstract

There are two variants of the classical multi-armed bandit (MAB) problem that have received considerable attention from machine learning researchers in recent years: contextual bandits and simple regret minimization. The contextual bandit problem is a generalization of MAB where the context is time varying. At every time step, in contextual bandits, the learner has access to side information that is predictive of the best arm. Simple regret minimization assumes that the learner only incurs regret after a pure exploration phase. In this work, we study simple regret minimization for contextual bandits. We present the Contextual-Gap algorithm and establish performance guarantees on the simple regret, i.e., the regret during the pure exploitation phase. Our experiments examine a novel application to adaptive sensor selection for magnetic field estimation in interplanetary spacecraft, and demonstrate considerable improvement over algorithms designed to minimize the cumulative regret.

1. Introduction

The multi-armed bandit (MAB) is a framework for sequential decision making where, at every time step, the learner selects (or "pulls") one of several possible actions (or "arms"), and receives a reward based on the selected action. The regret of the learner is the difference between the maximum possible reward and the reward resulting from the chosen action. In the classical MAB setting, the goal is to minimize the sum of all regrets, or *cumulative regret*, which naturally leads to an exploration/exploitation trade-off problem (Auer et al., 2002).

The contextual bandit problem extends the classical MAB setting, with the addition of time-varying side information, or *context*, made available at every time step. The best arm at every time step depends on the context, and intuitively the learner seeks to determine the best arm as a function of context. To date, work on contextual bandits has studied cumulative regret minimization, which is motivated by applications in health care, web advertisement recommendations and news article recommendations (Li et al., 2010).

In classical (non-contextual) MABs, the goal of the learner isn't always to minimize the cumulative regret. In some applications, there is a *pure exploration* phase during which the learning incurs no regret (i.e., no penalty for sub-optimal decisions), and performance is measured in terms of *simple regret*, which is the regret assessed at the end of the pure exploration phase(Gabillon et al., 2012; Jamieson & Nowak, 2014; Garivier & Kaufmann, 2016; Carpentier & Valko, 2015).

In this paper, we extend the idea of simple regret minimization to contextual bandits. In this setting, there is a pure exploration phase during which no regret is incurred, following by a *pure exploitation* phase during which regret is incurred, but there is no feedback so the learner cannot update its policy. To our knowledge, previous work has not addressed novel algorithms for this setting. Guan & Jiang (2018) provide simple regret guarantees for the policy of uniform sampling of arms in the i.i.d setting. The contextual bandit algorithm of Tekin & van der Schaar (2015) also has distinct exploration and exploitation phases, but unlike our setting, the agent has control over which phase it is in, i.e., when it wants to receive feedback. In the work of Hoffman et al. (2014); Soare et al. (2014); Libin et al. (2017); Xu et al. (2018) there is a single best arm even when contexts are observed (directly or indirectly). Our algorithm, Contextual-Gap, generalizes the idea of Bayes Gap (Hoffman et al., 2014) and UGapEb (Gabillon et al., 2012) to the contextual bandits setting.

We make the following contributions: 1. We formulate a novel problem: that of simple regret minimization for contextual bandits. 2. We develop an algorithm, Contextual-Gap, for this setting. 3. We present performance guarantees

[*] Equal contribution [1] Bing Ads, Microsoft AI & Research, USA [2] Microsoft Search, Assistant & Intelligence, Microsoft, USA [3] Department of Aerospace Engineering, University of Michigan Ann Arbor, USA [4] Climate and Space Engineering, University of Michigan Ann Arbor, USA [5] Department of EECS, University of Michigan Ann Arbor, USA. Correspondence to: Aniket Anand Deshmukh <aniketde@umich.edu>, Srinagesh Sharma <srinag@umich.edu>.

Proceedings of the 2^{nd} Exploration in Reinforcement Learning Workshop at the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

on the simple regret in the fixed budget framework. 4. We present experimental results for adaptive sensor selection in nano-satellites.

2. Motivation

Our work is motivated by autonomous systems that go through an initial training phase (the pure exploration phase) where they learn how to accomplish a task without being penalized for sub-optimal decisions, and then are deployed in an environment where they no longer receive feedback, but regret is incurred (the pure exploitation phase).

An example scenario arises in the problem of weak interplanetary magnetic field estimation using resource-constrained spacecraft known as nano-satellites or CubeSats. Spacecrafts produce spatially localized magnetic field noise due to large numbers of time-varying current paths in the spacecraft. Recent work in magnetic field noise minimization has focused on nano-satellites with multiple magnetic field sensors and adaptive methods for reducing spacecraft noise (Sheinker & Moldwin, 2016). At each time step, whenever a sensor is selected, the measurements go through a computationally expensive and power hungry calibration process (Kepko et al., 1996; Leinweber, 2012), which has to be repeated for every sensor at every timestep. Due to these constraints, it is required to select a single sensor at each time step.

Furthermore, the best sensor changes with time. This stems from the time-varying localization of noise in the spacecraft caused by operational events such as data transmission, spacecraft maneuvers, and power generation. This dynamic sensor selection problem is readily cast as a contextual bandit problem. The context is given by the spacecraft's telemetry which provides real-time measurements related to spacecraft operational events (Springmann & Cutler, 2012).

In this application, however, conventional contextual bandit algorithms are not applicable because feedback is not always available. Feedback requires knowledge of sensor noise, which in turn requires knowledge of the true magnetic field. Yet the true magnetic field is known only during certain portions of a spacecraft's orbit. Moreover, when the true magnetic field is known, there is no need to estimate the magnetic field in the first place! This suggests a learning scenario where the agent (the sensor scheduler) operates in two phases, one where it has feedback but incurs no regrets (because the field being estimated is known), and another where it does not receive feedback, but nonetheless needs to produce estimates. This is precisely the problem we study.

In the motivating problem, the exploration and exploitation occurs in phases, as the satellite moves into and out of regions where the true magnetic field is known. For simplicity, we address the problem in which the first T time steps belong to the exploration phase, and all subsequent time steps to the exploitation phase. Nonetheless, the algorithm we introduce can switch between phases indefinitely, and does not need to know in advance when a new phase is beginning.

3. Formal Setting

We denote the context space as $\mathcal{X} = \mathbb{R}^d$. Let $\{x_t\}_{t=1}^{\infty}$ denote the sequence of observed contexts. Let the total number of arms be A. For each x_t, the learner is required to choose an arm $a \in [A]$, where $[A] := \{1, 2, ..., A\}$.

For arm $a \in [A]$, let $f_a : \mathcal{X} \to \mathbb{R}$ be a function that maps context to expected reward when arm a is selected. Let a_t denote the arm selected at time t, and assume the reward at time t obeys $r_t := f_{a_t}(x_t) + \zeta_t$, where ζ_t is noise (described in more detail below). We assume that for each a, f_a belongs to a reproducing kernel Hilbert space (RKHS) defined on $\mathcal{X}$. The first T time steps belong to the *exploration phase* where the learner observes context x_t, chooses arm a_t and obtains reward r_t. The time steps after T belong to an *exploitation phase* where the learner observes context x_t, chooses arm a_t and earns an implicit reward r_t that is not returned to the learner.

For the theoretical results below, the following general probabilistic framework is adopted, following Abbasi-Yadkori et al. (2011) and Durand et al. (2018). We assume that ζ_t is a zero mean, ρ-conditionally sub-Gaussian random variable, i.e., ζ_t is such that for some $\rho > 0$ and $\forall \gamma \in \mathbb{R}$,

$$\mathbb{E}[e^{\gamma \zeta_t} | \mathcal{H}_{t-1}] \leq \exp\left(\frac{\gamma^2 \rho^2}{2}\right). \tag{1}$$

Here $\mathcal{H}_{t-1} = \{x_1, \ldots, x_{t-1}, \zeta_1, \ldots, \zeta_{t-1}\}$ is the history at time t (see supplementary material for additional details).

We also define the following terms. Let $D_{a,t}$ be the set of all time indices when arm a was selected up to time $t-1$ and set $N_{a,t} = |D_{a,t}|$. Let $X_{a,t}$ be the data matrix whose columns are $\{x_\tau\}_{\tau \in D_{a,t}}$ and similarly let $Y_{a,t}$ denote the column vector of rewards $\{r_\tau\}_{\tau \in D_{a,t}}$. Thus, $X_{a,t} \in \mathbb{R}^{d \times N_{a,t}}$ and $Y_{a,t} \in \mathbb{R}^{N_{a,t}}$.

3.1. Problem Statement

At every time step t, the learner observes context x_t. During the exploration phase $t \leq T$, the learner chooses a series of actions to explore and learn the mappings f_a from context to reward. During the exploitation phase $t > T$, the goal is to select the best arm as a function of context. We define the *simple regret* associated with choosing arm $a \in [A]$, given context x, as:

$$R_a(x) := f^*(x) - f_a(x), \tag{2}$$

where $f^*(x) := \max_{i \in [A]} f_i(x)$ is the expected reward for the best arm for context x. The learner aims to minimize the simple regret for $t > T$. To be more precise, let Ω be the fixed policy mapping context to arm during the exploitation phase. The goal is to determine policies for the exploration and exploitation phases such that for all $\epsilon > 0$ and $t > T$

$$\mathbb{P}(R_{\Omega(x_t)}(x_t) \geq \epsilon | x_t) \leq b_\epsilon(T),$$

where $b_\epsilon(T)$ is an expression that decreases to 0 as $T \to \infty$. The following section presents an algorithm to solve this problem.

4. Algorithm

We propose an algorithm that extends the Bayes Gap algorithm (Hoffman et al., 2014) to the contextual setting. Note that Bayes Gap itself is originally motivated from UGapEb (Gabillon et al., 2012). We describe how to get reward estimates $\hat{f}_{a,t}(x_t)$ and confidence estimates $s_{a,t}(x_t)$ in the supplementary material using kernel methods.

4.1. Contextual-Gap Algorithm

Algorithm 1 Contextual-Gap

```
Input: Number of arms $A$, Time Steps $T$, parameter $\beta$,
regularization parameter $\lambda$, burn-in phase constant $N_\lambda$.
// Exploration Phase I: Burn-in Period //
for $t = 1, ..., AN_\lambda$ do
    Observe $x_t$, choose $a_t = t \mod A$ and receive $r_t \in \mathbb{R}$
end for
//Exploration Phase II: Contextual-Gap Policy //
for $t = AN_\lambda + 1, \ldots, T$ do
    Observe context $x_t$
    Learn reward estimators $\hat{f}_{a,t}(x_t)$ and confidence interval $s_{a,t}(x_t)$ based on history
    $U_{a,t}(x_t) = \hat{f}_{a,t}(x_t) + \frac{s_{a,t}(x_t)}{2}$
    $L_{a,t}(x_t) = \hat{f}_{a,t}(x_t) - \frac{s_{a,t}(x_t)}{2}$
    $B_{a,t}(x_t) = \max_{i \neq a} U_{i,t}(x_t) - L_{a,t}(x_t)$
    $J_t(x_t) = \arg\min_a B_{a,t}(x_t)$
    $j_t(x_t) = \arg\max_{a \neq J_t(x_t)} U_{a,t}(x_t)$
    Choose $a_t = \arg\max_{a \in \{j_t(x_t), J_t(x_t)\}} s_{a,t}(x_t)$
    Receive reward $r_t \in \mathbb{R}$
end for
// Exploitation Phase //
for $t > T$ do
    Observe context $x_t$.
    for $\tau = AN_\lambda + 1, \ldots, T$ do
        Evaluate and collect $J_\tau(x_t), B_{J_\tau(x_t)}(x_t)$
    end for
    $\iota = \arg\min_{AN_\lambda + 1 \leq \tau \leq T} B_{J_\tau(x_t),t}(x_t)$
    Choose $\Omega(x_t) = J_\iota(x_t)$.
end for
```

During the exploration phase, the Contextual-Gap algorithm proceeds as follows. First, the algorithm has a burn-in period where it cycles through the arms (ignoring context) and pulls each one N_λ times. Following this burn-in phase, when the algorithm is presented with context x at time $t \leq T$, the algorithm identifies two candidate arms, $J_t(x)$ and $j_t(x)$, as follows. For each arm a the *contextual gap* is defined as $B_{a,t}(x) := \max_{i \neq a} U_{i,t}(x) - L_{a,t}(x)$. $J_t(x)$ is the arm that *minimizes* $B_{a,t}(x)$ and $j_t(x)$ is the arm (excluding $J_t(x)$) whose upper confidence bound is maximized. Among these two candidates, the one with the widest confidence interval is selected. Quantity $B_{a,t}(x)$ is used to bound the the simple regret for corresponding arm a and is the basis of definition of the best arm $J_t(x)$. We use $j_t(x) = \arg\max_{a \neq J_t(x)} U_{a,t}(x)$ as the second candidate because optimistically $j_t(x)$ has a chance to be the best arm and it may give more information about how bad the choice of $J_t(x)$ could be.

In the exploitation phase, for a given context x, the contextual gap for all time steps in the exploration phase are evaluated. The arm with the smallest gap over the entire exploration phase for the given context x is chosen as the best arm associated with context x. Because there is no feedback during the exploitation phase, the algorithm moves to the next exploitation step without modification to the learning history. The exact description is presented in Algorithm 1. We give more intuition about the algorithm in the Section 3 of supplementary material.

During the exploitation phase, looking back at all history may be computationally prohibitive. Thus, in practice, we just select the best arm as $J_T(x_t), \forall t > T$. As described in the experimental section, this works well in practice. Theoretically, N_λ has to be bigger than a certain number defined in Theorem 4.1, but for experimental results we keep $N_\lambda = 1$.

4.2. Learning Theoretic Analysis

We provide high probability simple regret bounds in the non-i.i.d setting described in section 3. The algorithm operates under very general assumptions on the context and the kernel function that are stated as follows:

A I $\{\mathcal{X}_t\}_{t \geq 1} \subset \mathbb{R}^d$, is a random process on compact space endowed with a finite positive Borel measure.

A II Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded by a constant L, the canonical feature map $\phi : \mathcal{X} \to \mathcal{H}$ of k is a continuous function, and $\mathcal{H}$ is separable.

For the confidence interval to be useful, it needs to shrink to zero with high probability over the feature space as each arm is pulled more and more. This requires the smallest non-zero eigenvalue of the sample covariance matrix of the data for each arm to be lower bounded by a certain value. We make an assumption to allow for such a lower bound.

Denote $\mathbb{E}_{t-1}[\cdot] := \mathbb{E}[\cdot|x_1, x_2, \ldots, x_{t-1}]$ and by $\lambda_r(A)$ the r^{th} largest eigenvalue of a compact self adjoint operator A. For a context x, the operator $\phi(x)\phi(x)^T : \mathcal{H} \rightarrow \mathcal{H}$ is a compact self-adjoint operator.

A III There exists a subspace of dimension d^* with projection P, and a constant $\lambda_x > 0$, such that $\forall t$, $\lambda_r(P^T\mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T]P) > \lambda_x$ for $r \leq d^*$ and $\lambda_r((I-P)^T\mathbb{E}_{t-1}[\phi(x_t)\phi(x_t)^T](I-P)) = 0, \forall r > d^*$.

The upper bound depends on a context-based hardness quantity defined for each arm a (similar to Hoffman et al. (2014)) as

$$H_{a,\epsilon}(x) = \max(\frac{1}{2}(\Delta_a(x) + \epsilon), \epsilon). \tag{3}$$

where $\Delta_a(x) := |\max_{i \neq a} f_i(x) - f_a(x)|$ is the the gap quantity. Denote its lowest value as $H_{a,\epsilon} := \inf_{x \in \mathcal{X}} H_{a,\epsilon}(x)$. Let total hardness be defined as $H_\epsilon := \sum_{a \in [A]} H_{a,\epsilon}^{-2}$ (Note that $H_\epsilon \leq \frac{A}{\epsilon^2}$). The recommended arm after time $t \geq T$ is defined as

$$\Omega(x) = J_{\arg\min_{AN_\lambda + 1 \leq \tau \leq T} B_{J_\tau(x_t),t}(x_t)}(x_t)$$

from Algorithm 1. Using the assumptions and the hardness quantities, we upper bound the simple regret as follows:

Theorem 4.1 *Consider a contextual bandit problem as defined in Section 3 with assumptions **A I-A III**. For* $0 < \delta \leq \frac{1}{8}$, $\epsilon > 0$ *and* $N_\lambda := \max\left(\frac{2(1-\lambda)}{\lambda_x}, d^*, \frac{256}{\lambda_x^2}\log(\frac{128\tilde{d}}{\lambda_x^2\delta})\right)$, *let*

$$\beta = \sqrt{\frac{\lambda_x(T - N_\lambda(A-1)) + 2A\lambda}{16C_1^2 H_\epsilon}} - \frac{C_2}{C_1}. \tag{4}$$

for constants C_1 *and* C_2. *We have for all* $t > T$ *and* $\epsilon > 0$,

$$\mathbb{P}(R_{\Omega(x_t)}(x_t) < \epsilon | x_t) \geq 1 - A(T - AN_\lambda)e^{-\beta^2} - A\delta. \tag{5}$$

The detailed analysis and proofs are provided in the supplementary material.

5. Experimental Results

We present results from a lab generated non-i.i.d spacecraft magnetic field as described in Section 2. We also present results on synthetic data in the supplementary material. We present average simple regret comparisons of the Contextual-Gap algorithm against five baselines: Uniform sampling, Epsilon Greedy, kernel-UCB ((Valko et al., 2013)), Kernel-UCB-Mod (Kernel UCB for exploration but best estimated reward for exploitation, Kernel Thompson Sampling (Chowdhury & Gopalan, 2017)

For all the algorithms, we use the Gaussian kernel and tune the bandwidth of the kernel, and the regularization parameter (more details in supplementary material). The tuned parameters were used with the evaluation datasets to generate the plots. The code is available online to reproduce all results [1].

5.1. Experimental Spacecraft Magnetic Field Dataset

We present the experimental setup and results associated with a lab generated, realistic spacecraft magnetic field dataset with *non-i.i.d contexts*. In spacecraft magnetic field data, we are interested in identifying the least noisy sensor for every time step (see Section 2).

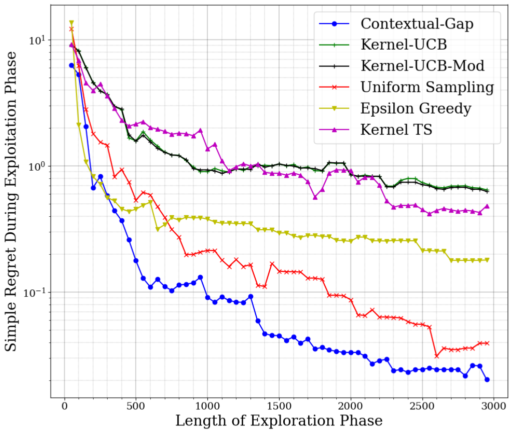

Figure 1: Average Simple Regret Evaluation on Spacecraft Magnetic Field Dataset

The dataset was generated with contexts x_t consisting of measured variables associated with the electrical behavior of the GRIFEX spacecraft (Norton et al., 2012; Cutler et al., 2015), and reward is the negative of the magnitude of the sensor noise measured at every time step. Data were collected using 3 sensors (arms) and the true magnetic field was computed using models of the earth's magnetic field.

Figure 1 shows the simple regret minimization curves for the spacecraft data-set and even in this case Contextual-Gap converges faster compared to other algorithms.

6. Conclusion

In this work, we present a novel problem: that of simple regret minimization in the contextual bandit setting. We propose the Contextual-Gap algorithm, give a regret bound for the simple regret, and show empirical results on three multiclass datasets and one lab-based spacecraft magnetometer dataset. It can be seen that in this scenario persistent and efficient exploration of the best and second best arms with the Contextual-Gap algorithm provides improved results compared against algorithms designed to optimize cumulative regret.

[1] The code to reproduce our results is available at `https://github.com/aniketde/ContextualGap`

Acknowledgements

Clayton Scott and Aniket Anand Deshmukh were supported in part by NSF Grants 1422157 and 1838179. Mark Moldwin, James W. Cutler and Srinagesh Sharma were supported in part by NASA SSTP Grant NNX16AT35A.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Carpentier, A. and Valko, M. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pp. 1133–1141, 2015.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853, 2017.

Cutler, J. W., Lacy, C., Rose, T., Kang, S.-h., Rider, D., and Norton, C. An update on the GRIFEX mission. *Cubesat Developer's Workshop*, 2015.

Durand, A., Maillard, O.-A., and Pineau, J. Streaming kernel regression with provably adaptive mean, variance, and regularization. *Journal of Machine Learning Research*, 19(August), 2018.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pp. 3212–3220, 2012.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027, 2016.

Guan, M. Y. and Jiang, H. Nonparametric stochastic contextual bandits. In *The 32nd AAAI Conference on Artificial Intelligence*, 2018.

Hoffman, M., Shahriari, B., and Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, pp. 365–374, 2014.

Jamieson, K. and Nowak, R. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pp. 1–6. IEEE, 2014.

Kepko, E. L., Khurana, K. K., Kivelson, M. G., Elphic, R. C., and Russell, C. T. Accurate determination of magnetic field gradients from four point vector measurements. I. use of natural constraints on vector data obtained from a single spinning spacecraft. *IEEE Transactions on Magnetics*, 32(2):377–385, 1996.

Leinweber, H. K. *In-flight calibration of space-borne magnetometers*. PhD thesis, Graz University of Technology, 2012.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Libin, P., Verstraeten, T., Roijers, D. M., Grujic, J., Theys, K., Lemey, P., and Nowé, A. Bayesian best-arm identification for selecting influenza mitigation strategies. *arXiv preprint arXiv:1711.06299*, 2017.

Norton, C. D., Pasciuto, M. P., Pingree, P., Chien, S., and Rider, D. Spaceborne flight validation of NASA ESTO technologies. In *Geoscience and Remote Sensing Symposium (IGARSS), 2012 IEEE International*, pp. 5650–5653. IEEE, 2012.

Sheinker, A. and Moldwin, M. B. Adaptive interference cancelation using a pair of magnetometers. *IEEE Transactions on Aerospace and Electronic Systems*, 52(1):307–318, 2016.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pp. 828–836, 2014.

Springmann, J. C. and Cutler, J. W. Attitude-independent magnetometer calibration with time-varying bias. *Journal of Guidance, Control, and Dynamics*, 35(4):1080–1088, 2012.

Tekin, C. and van der Schaar, M. Releaf: An algorithm for learning and exploiting relevance. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):716–727, 2015.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. Finite-time analysis of kernelised contextual bandits. In *Uncertainty in Artificial Intelligence*, pp. 654. Citeseer, 2013.

Xu, L., Honda, J., and Sugiyama, M. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 843–851, 2018.