Multi-Armed Bandits with Correlated Arms

Samarth Gupta [1] **Shreyas Chaudhari** [1] **Gauri Joshi** [1] **Osman Yagan** [1]

Abstract

We consider a novel multi-armed bandit framework where the rewards obtained by pulling different arms are correlated. We develop a unified approach to leverage these reward correlations and present fundamental generalizations of classic bandit algorithms to the correlated setting. Regret analysis of C-UCB and C-TS (the correlated bandit versions of Upper-confidence-bound and Thompson sampling) reveals that the algorithms end up pulling certain sub-optimal arms, termed as *non-competitive*, only $\mathrm{O}(1)$ times, as opposed to the $\mathrm{O}(\log T)$ pulls required by classic bandit algorithms such as UCB, TS etc. We validate the proposed algorithms via experiments on the MovieLens dataset, and show significant improvement over classical bandit algorithms.

The full version of the paper with additional examples, results, proofs and experiments is available at: https://www.andrew.cmu.edu/user/samarthg/MABCorr.pdf

1. Introduction

Classical Multi-armed Bandits. In the classical multi-armed bandit (MAB) problem, there are K possible actions, referred to as arms, with each arm having an unknown reward distribution. At each round t, we need to choose an arm $k_t \in \mathcal{K}$ and we receive a random reward R_t drawn from the reward distribution of arm k_t. The goal is to maximize the cumulative reward over a horizon of T time slots. In order to maximize cumulative reward, it is important to balance the exploration-exploitation trade-off, i.e., learning the mean reward of each arm while trying to make sure that the arm with the highest mean reward is played as many times as possible. The problem has been extensively studied (Lai and Robbins, 1985) and has proven to be useful in numerous applications including A/B Testing (White, 2012), ad placement, recommendation systems, clinical trials (Villar et al., 2015), system testing (Tekin and Turgay, 2017).

[1]Carnegie Mellon University, Pittsburgh, USA. Correspondence to: Samarth Gupta <samarthg@andrew.cmu.edu>.

Preliminary work. Under review by the Theoretical foundations of reinforcement learning workshop at International Conference on Machine Learning (ICML). Do not distribute.

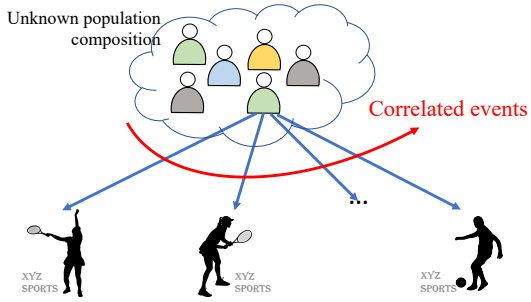

Figure 1: A user's ratings for different versions of the same ad are correlated. If a user likes the first ad, there is a good chance that they will also like the second since it is also related to tennis. However, since the population composition (the fraction of people liking the first/second or the last version) is unknown, it is not clear what is a good global recommendation for the population.

Consider the application of advertisement selection, where a company needs to decide the which ad-version to display to it's users in order to maximize the user engagement over the course of their ad-campaign (See Figure 1). The response of a user corresponding to two different ad-versions is likely to be correlated in practice, for instance, a user reacting positively (by clicking, ordering, etc.) to the first version of the ad with a girl playing tennis might be more likely to click the second version as it is also related to tennis; of course one can construct examples where there is negative correlation between click events to different ads. The model we study in this paper explicitly captures these correlations, something that has not been studied previously. Unlike contextual bandits (Zhou, 2015), we do not observe the context (age/occupational/income) features of the user and do not focus on providing personalized recommendation. Instead our goal is to provide global recommendations to a population whose demographics is unknown. Unlike *structured bandits* (Combes et al., 2017), we do not assume that the mean rewards are functions of a hidden context θ.

Model overview. We capture correlation between rewards from different arms in the form of *pseudo-rewards*, which provide upper bounds on the conditional expectation of rewards. In the context of displaying ad versions, pseudo-rewards represent an upper bound on *the probability that user likes version B of the ad if it liked/disliked version A*.

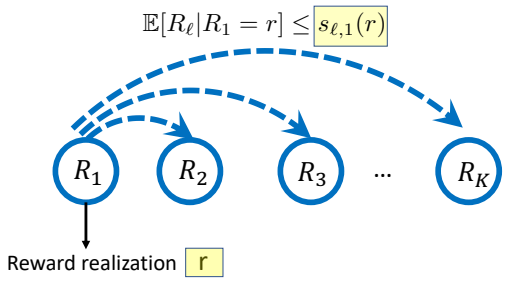

Figure 2: Upon observing a reward r from an arm k, pseudo-rewards $s_{\ell,k}(r)$, give us an upper bound on the conditional expectation of the reward from arm ℓ given that we observed reward r from arm k. These pseudo-rewards models the correlation in rewards corresponding to different arms.

Figure 2 presents an illustration of our correlation model, where the pseudo-rewards, denoted by $s_{\ell,k}(r)$, provide an upper bound on the reward that we could have received from arm ℓ given that pulling arm k led to a reward of r. In practice, pseudo-rewards can be obtained via expert/domain knowledge (for example, common ingredients in two drugs that are being considered to treat an ailment) or controlled surveys (for example, beta-testing users who are asked to rate different versions of an ad). A key advantage of our framework is that pseudo-rewards are just upper bounds on the conditional expected rewards and can be arbitrarily loose. This also makes the proposed framework and algorithm directly usable in practice – if some pseudo-rewards are unknown due to lack of domain knowledge/data, they can simply be replaced by the maximum possible reward entries, which serves a natural upper bound.

Aside from the novel correlated bandit model, we propose an approach that fundamentally generalizes any classical bandit algorithm to the correlated-bandit setting and provide regret bounds for them through a unified regret analysis.

2. Problem Formulation

Consider a multi-armed bandit setting with K arms $\{1, 2, \dots K\}$. At each round t, a user enters the system and we need to decide an arm k_t to display to the user. Upon displaying arm k_t, we receive a random reward $R_{k_t} \in [0, B]$. Our goal is to maximize the cumulative reward over time. The expected reward (over the population of users) of arm k, is denoted by μ_k. If we knew the arm with highest mean, i.e., $k^* = \arg\max_{k \in \mathcal{K}} \mu_k$ beforehand, then we would always pull arm k^*. We now define the cumulative regret, minimizing which is equivalent to maximizing cumulative reward:

$$\mathbb{E}\left[Reg(T)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mu_{k_t} - \mu_{k^*}\right] = \sum_{k \neq k^*} \mathbb{E}\left[n_k(T)\right] \Delta_k. \tag{1}$$

r	$s_{2,1}(r)$
0	0.7
1	0.4

r	$s_{1,2}(r)$
0	0.8
1	0.5

(a)	$R_1 = 0$	$R_1 = 1$
$R_2 = 0$	0.2	0.4
$R_2 = 1$	0.2	0.2

(b)	$R_1 = 0$	$R_1 = 1$
$R_2 = 0$	0.2	0.3
$R_2 = 1$	0.4	0.1

Table 1: The top row shows the pseudo-rewards of arms 1 and 2, i.e., upper bounds on the conditional expected rewards (which are known to the player). The bottom row depicts two possible joint probability distribution (unknown to the player). Under distribution (a), Arm 1 is optimal whereas Arm 2 is optimal for distribution (b).

Here, $n_k(T)$ denotes the number of times a sub-optimal arm is pulled till round T and Δ_k denotes the *sub-optimality gap* of arm k, i.e., $\Delta_k = \mu_{k^*} - \mu_k$. The standard multi-armed bandit setting does not explicitly account for known correlations between rewards. In many practical scenarios, rewards corresponding to different arms are known to be correlated. Motivated by this, we consider a setup where the conditional distribution of the reward from arm ℓ given reward from arm k is not equal to the probability distribution of the reward from arm ℓ, i.e., $f_{R_\ell|R_k}(r_\ell|r_k) \neq f_{R_\ell}(r_\ell)$, with $f_{R_\ell}(r_\ell)$ denoting the probability distribution function of the reward from arm ℓ. Consequently, due to such correlations, we have $\mathbb{E}\left[R_\ell|R_k\right] \neq \mathbb{E}\left[R_\ell\right]$. We model this correlation explicitly by the knowledge of *pseudo-rewards* that constitute an upper bound on conditional expected rewards.

Definition 1 (Pseudo-Reward)**.** *Suppose we pull arm k and observe reward r, then the pseudo-reward of arm ℓ with respect to arm k, denoted by $s_{\ell,k}(r)$, is an upper bound on the conditional expected reward of arm ℓ, i.e.,*

$$\mathbb{E}[R_\ell|R_k = r] \leq s_{\ell,k}(r). \tag{2}$$

These pseudo-rewards can be learned from historical data or through *offline* surveys in which users are presented with *all* K arms allowing us to sample $R_1, \dots, R_K$ jointly. For example in Table 1, we can look at all users who obtained 0 reward from Arm 1 and calculate the corresponding average reward $\hat{\mu}_{2,1}(0)$ from Arm 2. If the training data is large, this value is close to $\mathbb{E}\left[R_2|R_1 = 0\right]$ and can be used directly as $s_{2,1}(0)$. Alternately, we can set $s_{2,1}(0) = \hat{\mu}_{2,1}(0) + \hat{\sigma}_{2,1}(0)$, where $\hat{\sigma}_{2,1}(0)$ is the empirical standard deviation of the conditional reward of arm 2, which is added as a safety buffer. In the absence of joint samples, the pseudo-rewards can be set to the maximum possible reward of that arm.

Remark 1. *When all pseudo-reward entries are unknown, then all pseudo-reward entries can be filled with maximum possible reward for each arm. In such a case, the proposed* C-BANDIT *algorithm reduces to the underlying classic* BANDIT *(for e.g., UCB, TS etc.) algorithm.*

Comparison with parametric models As mentioned in

Section 1, a seemingly related model is the structured bandits model (Lattimore and Munos, 2014). Structured bandits is a class of problems that cover linear bandits (Abbasi-Yadkori et al., 2011), generalized linear bandits (Filippi et al., 2010), Lipschitz bandits (Magureanu et al., 2014), global bandits (Atan et al., 2015), regional bandits (Wang et al., 2018) etc. In the structured bandits setup, mean rewards corresponding to different arms are related to one another through a hidden parameter θ. The underlying value of θ is fixed and the mean reward mappings $\theta \rightarrow \mu_k(\theta)$ are known. Similarly, (Pandey et al., 2007) studies a dependent armed bandit problem, that also has mean rewards corresponding to different arms related to one another. All of these models are fundamentally different from the problem setting considered in this paper. In this work we explicitly model the correlations in the rewards of a user corresponding to different arms. While, mean rewards are related to each other in structured bandits and (Pandey et al., 2007), the reward realizations are not necessarily correlated. Another key difference is that the model studied here is non-parametric in the sense that there is no hidden feature space as is the case in structured bandits and (Pandey et al., 2007).

3. Proposed C-BANDIT Algorithms

We now propose an approach that extends any classical multi-armed bandit algorithm (such as UCB, Thompson Sampling, KL-UCB) to the correlated MAB setting by making use of the pseudo-rewards.

Definition 2 (Empirical and Expected Pseudo-Reward)**.** *If arm k is pulled $n_k(t)$ times in t rounds, these $n_k(t)$ reward realizations can be used to construct the empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ for each arm ℓ with respect to arm k, which is defined as follows.*

$$\hat{\phi}_{\ell,k}(t) \triangleq \frac{\sum_{\tau=1}^{t} \mathbb{1}_{k_\tau=k}\, s_{\ell,k}(r_\tau)}{n_k(t)}, \quad \ell \in \{1,\ldots,K\}\setminus\{k\}.$$

w.l.o.g., we set $\hat{\phi}_{k,k} = \hat{\mu}_k$. As $n_k(t) \rightarrow \infty$, $\hat{\phi}_{\ell,k}(t) \rightarrow \phi_{\ell,k} \triangleq \mathbb{E}[s_{\ell,k}(R)]$, the expected pseudo-reward of arm ℓ with respect to arm k.

Note that the empirical pseudo-reward $\hat{\phi}_{\ell,k}(t)$ is defined with respect to arm k and they provide an estimate on the upper bound of the mean of arm ℓ, i.e, μ_ℓ, through only the reward samples obtained from arm k. In each round, the algorithm performs the following steps:

1. **Identify the set of significant arms $\mathcal{S}_t$**: $\mathcal{S}_t = \{k : n_k(t) > \frac{t}{K}\}$. Furthermore, define $k^{\text{emp}}(t) = \arg\max_{k\in\mathcal{S}_t} \hat{\mu}_k(t)$.
2. **Identify empirically competitive arms $\mathcal{A}_t$**: Identify the set $\mathcal{A}_t$ of arms that are empirically competitive with respect to the set $\mathcal{S}_t$, i.e.,
$$\mathcal{A}_t = \{k \in \mathcal{K} : \hat{\mu}_{k^{\text{emp}}}(t) \leq \min_{\ell\in\mathcal{S}_t} \hat{\phi}_{k,\ell}(t)\}.$$
3. **Choose an arm from $\{\mathcal{A}_t \cup k^{emp}(t)\}$** using a BANDIT algorithm (eg. UCB, Thompson sampling, KL-UCB etc.): For instance, the C-UCB pulls the arm
$$k_{t+1} = \arg\max_{k\in\{\mathcal{A}_t\cup k^{emp}\}} I_{k,t},$$
where $I_{k,t} = \hat{\mu}_k(t) + B\sqrt{\frac{2\log(t)}{n_k(t)}}$, the UCB index (Auer et al., 2002). Similarly, C-TS pulls the arm $k_{t+1} = \arg\max_{k\in\{\mathcal{A}_t\cup k^{emp}\}} S_{k,t}$, where $S_{k,t} \sim \mathcal{N}\left(\hat{\mu}_k(t), \frac{\beta B}{n_k(t)}\right)$, the sample obtained from the posterior distribution of μ_k (Agrawal and Goyal, 2013),
4. Update the empirical pseudo-rewards $\hat{\phi}_{\ell,k_t}(t+1)$ for all ℓ, the empirical reward for arm k_{t+1}.

At each round t, through samples of arms in $\mathcal{S}_t$, we eliminate some arms which are not empirically competitive at round t and do not consider them in step 3 of the algorithm. By performing this elimination at each round, we reduce the amount of exploration in our algorithms as some arms may be viewed as sub-optimal for round t based on just the samples of arms in $\mathcal{S}_t$. Note that this elimination is done only for a single round, i.e., an arm that is not empirically competitive at round t may become empirically competitive in future rounds. A key strength of our approach is that we can use any standard bandit algorithm (UCB, TS, KL-UCB (Garivier and Cappé, 2011), Bayes-UCB (Kaufmann et al., 2012) etc.) in step 3 of the above algorithm.

4. Regret Analysis and Bounds

In order to bound $\mathbb{E}[Reg(T)]$, we can analyze the expected number of times sub-optimal arms are pulled, that is, $\mathbb{E}[n_k(T)]$, for all $k \neq k^*$. Theorem 1 and Theorem 2 below show that $\mathbb{E}[n_k(T)]$ scales as $O(1)$ and $O(\log T)$ for non-competitive and competitive arms respectively.

Definition 3 (Non-Competitive and Competitive arms)**.** *An arm ℓ is said to be non-competitive if the expected reward of optimal arm k^* is larger than the expected pseudo-reward of arm ℓ with respect to the optimal arm k^*, i.e, if, the pseudo-gap of arm ℓ, $\tilde{\Delta}_{\ell,k^*} \triangleq \mu_{k^*} - \phi_{\ell,k^*} > 0$. Similarly, an arm ℓ is said to be competitive if $\tilde{\Delta}_{\ell,k^*} = \mu_{k^*} - \phi_{\ell,k^*} <= 0$. The unique best arm k^* has $\tilde{\Delta}_{k^*,k^*} = \mu_{k^*} - \phi_{k^*,k^*} = 0$ and is counted in the set of competitive arms.*

Theorem 1. *The expected number of times a non-competitive arm with pseudo-gap $\tilde{\Delta}_{k,k^*}$ is pulled by C-UCB is upper bounded as*

$$\mathbb{E}[n_k(T)] \leq Kt_0 + K^2 \sum_{t=Kt_0}^{T} 3\left(\frac{t}{K}\right)^{-2} + \sum_{t=1}^{T} t^{-3}, \tag{3}$$

$$= \mathrm{O}(1), \tag{4}$$

$$\textit{where, } t_0 = \inf\left\{\tau \geq 2 : \Delta_{min}, \tilde{\Delta}_{k,k^*} \geq 4\sqrt{\frac{K\log\tau}{\tau}}\right\}. \tag{5}$$

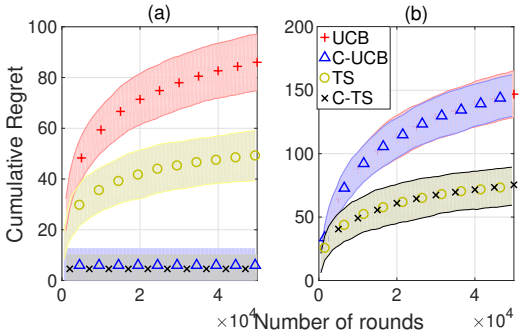

Figure 3: Cumulative regret for UCB, C-UCB and C-TS corresponding to the problem shown in Table 1. For the setting (a) in Table 1, Arm 1 is optimal and Arm 2 is non-competitive, in setting (b) of Table 1 Arm 2 is optimal while Arm 1 is competitive.

Theorem 2. *The expected number of times a competitive arm is pulled by C-UCB algorithm is upper bounded as*

$$\mathbb{E}\left[n_k(T)\right] \leq 8\frac{\log(T)}{\Delta_k^2} + \left(1+\frac{\pi^2}{3}\right) + \sum_{t=1}^{T} t\exp\left(-\frac{t\Delta_{min}^2}{2K}\right),$$
$$= \mathrm{O}(\log T) \quad \textit{where } \Delta_{min} = \min_k \Delta_k > 0. \tag{6}$$

Substituting the bounds on $\mathbb{E}\left[n_k(T)\right]$ derived in Theorem 1,2 into (1), we get the upper bound on expected regret as

$$\mathbb{E}\left[Reg(t)\right] \leq (C-1)\cdot \mathrm{O}(\log T) + (K-C)\cdot \mathrm{O}(1).$$

Reduction in effective number of arms. Since the distribution of reward of each arm is unknown, the C-UCB and C-TS algorithms initially do not know which arms are competitive. Even then, Theorem 1 shows that C-UCB makes sure that *non-competitive* arms are pulled only $\mathrm{O}(1)$ times. Due to this, only the competitive sub-optimal arms are pulled $\mathrm{O}(\log T)$ times. Moreover, the pre-log terms in the upper bound of UCB and C-UCB (and correspondingly TS and C-TS) for these arms is the same. In this sense, our C-BANDIT approach reduces a K-armed bandit problem to a C-armed bandit problem. Only $C-1 \leq K-1$ arms are pulled $\mathrm{O}(\log T)$ times, while the algorithm stops pulling other arms after $\mathrm{O}(1)$ rounds. When $C=1$, i.e., all sub-optimal arms are non-competitive, our proposed C-UCB and C-TS algorithms achieve $\mathrm{O}(1)$ regret, see Figure 3. [1]

5. Experiments

We use the MOVIELENS dataset (Harper and Konstan, 2015) to perform our experiments, which contains a total of 1M ratings for 3883 movies (split into 18 different genres) by a 6040 unique users on scale of 1 to 5. We split this dataset randomly into two halves, train and test dataset. We consider the goal of recommending the best of 18 possible movie genres to a population with unknown demographic. We use the training dataset to learn the pseudo-reward entries. The pseudo-reward entry $s_{\ell,k}(r)$ is evaluated by taking the empirical average of the ratings of genre ℓ that are rated by the users who rated genre k as r. To capture the fact that it might not be possible in practice to fill all pseudo-reward entries, we randomly remove a fraction p of the pseudo-reward entries. The removed pseudo-reward entries are replaced by the maximum possible rating, i.e., 5 (as that gives a natural upper bound on the conditional mean reward). Using these pseudo-rewards, we evaluate our proposed algorithms on the test data. Our experimental results for this setting are shown in Figure 4, with the fraction of removed pseudo-reward entries $p = 0.25, 0.50$ and 0.70. The proposed C-UCB and C-TS algorithms stop pulling some of the 18 arms within finite time and thus significantly outperform UCB in all three settings.

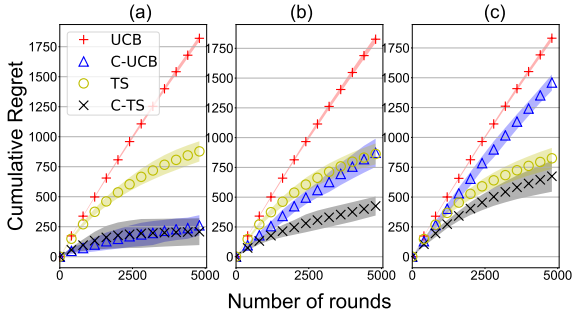

Figure 4: Cumulative regret for UCB, C-UCB and C-TS for the application of recommending the best genre (out of 18 genres) in the Movielens dataset, where p fraction of the pseudo-entries are replaced with maximum reward $i.e.,5$. In $(a), p = 0.25$, for $(b), p = 0.50$ and $p = 0.70$ in (c). The value of C is 4,11 and 13 in (a), (b) and (c) respectively.

In our plots, all data points are averaged over 100 runs. The shaded area represents error bars with one standard deviation.

6. Conclusion

This work studies the problem of regret minimization in a novel correlated Multi-Armed Bandit setting. The proposed algorithmic approach allows extension of any classical bandit algorithm to the correlated bandit setting and they significantly outperform classical bandit algorithms both theoretically and empirically. Open future directions include the design of best-arm identification algorithms for the correlated bandit setting.

[1] The rigorous proofs for C-TS showing $\mathrm{O}(\log T)$ pulls for competitive arms and $\mathrm{O}(1)$ pulls for non-competitive arms, and unified proof technique for all C-Bandit algorithms is available in the full version of the paper.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, pages 99–107, 2013.

Onur Atan, Cem Tekin, and Mihaela van der Schaar. Global multi-armed bandits with hölder continuity. In *AISTATS*, 2015.

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

Richard Combes, Stefan Magureanu, and Alexandre Proutière. Minimal exploration in structured stochastic bandits. In *NIPS*, 2017.

Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 359–376, 2011.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5, 4, Article 19, 2015.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Artificial intelligence and statistics*, pages 592–600, 2012.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Tor Lattimore and Rémi Munos. Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558, 2014.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. *arXiv preprint arXiv:1405.4758*, 2014.

Sandeep Pandey, Deepayan Chakrabarti, and Deepak Agarwal. Multi-armed bandit problems with dependent arms. In *Proceedings of the International Conference on Machine Learning*, pages 721–728, 2007.

Cem Tekin and Eralp Turgay. Multi-objective contextual multi-armed bandit problem with a dominant objective. *arXiv preprint arXiv:1708.05655*, 2017.

Sofía S Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199, 2015.

Zhiyang Wang, Ruida Zhou, and Cong Shen. Regional multi-armed bandits. In *AISTATS*, 2018.

John White. *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.", 2012.

Li Zhou. A survey on contextual multi-armed bandits. *CoRR*, abs/1508.03326, 2015. URL `http://arxiv.org/abs/1508.03326`.