Stay With Me: Lifetime Maximization Through Heteroscedastic Linear Bandits With Reneging

Ping-Chun Hsieh [*1] **Xi Liu** [*1] **Anirban Bhattacharya** [2] **P. R. Kumar** [1]

Abstract

Sequential decision making for lifetime maximization is a critical problem in many real-world applications, such as medical treatment and portfolio selection. In these applications, a "reneging" phenomenon, where participants may disengage from future interactions after observing an unsatisfiable outcome, is rather prevalent. To address the above issue, this paper proposes a model of heteroscedastic linear bandits with reneging, which allows each participant to have a distinct "satisfaction level," with any interaction outcome falling short of that level resulting in that participant reneging. Moreover, it allows the variance of the outcome to be context-dependent. Based on this model, we develop a UCB-type policy, namely HR-UCB, and prove that it achieves $\mathcal{O}\big(\sqrt{T(\log(T))^3}\big)$ regret. Finally, we validate the performance of HR-UCB via simulations.

1. Introduction

Sequential decision problems commonly arise in a large number of real-world applications. To name a few, in treatment to extend the life of people with terminal illnesses, doctors are required to make decisions on which treatments are used for patients periodically. In portfolio selection, fund managers need to decide which portfolios are recommended to their customers every time. In cloud computing services, the cloud platform has to determine the resources allocated to customers given specific requirements of their programs. Multi-armed Bandits (MAB) (Auer et al., 2002) and one of its most famous variants "contextual bandits" (Abbasi-Yadkori et al., 2011) have been extensively used to model such problems. In the modeling, available choices are referred to as "arms" and a decision is regarded as a "pull" of the corresponding arm. The decision is evaluated through rewards that depend on the goal of the interaction.

In the aforementioned applications and services, a phenomenon that participants may disengage from future interactions commonly exist. Such behavior is referred to as "churn", "unsubscribe" or "reneging" in literature (Liu et al., 2018). For instance, patients fail to survive the illnesses or are unable to take more treatments due to the deterioration of physical condition (McHugh et al., 2015). In portfolio selection, fund managers earn money from customer enrollment of the selection service. The return of the selection may turn out to be loss and thus the customer loses trust to the manager and stops using the service (Huo & Fu, 2017). Similarly, in the cloud computing services, the customer may feel the resource is not well allocated and leads to an unsatisfied throughput, thus switching to another service provider (Ding et al., 2013). In other words, the participant [1] of the interaction has a "lifetime" that can be defined as the total number interactions between the participant and a service provider until reneging. The larger the number is, the "longer" participant stay with the provider. Customer lifetime has been recognized as a critical metric to evaluate the success of many applications such as all above application as well as the e-commerce (Theocharous et al., 2015). Moreover, as well known, the acquisition cost for a new customer is much higher than an existing customer (Liu et al., 2018). Therefore, within the applications and services, one particular vital goal is to maximize the lifetime of customers. Our focus in this paper is to learn an optimal decision policy that maximizes the lifetime of participants in interactions.

We consider reneging behavior based on two observations. First, in all above scenarios, the decision maker is usually able to observe the outcome of following their suggestion, e.g., the physical condition of the patients after the treatment, the money earned from purchasing the suggested portfolio in the account, and the throughput rate of running the programs. Second, we observe that the participants in those applications are willing to reveal their satisfaction level to

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, USA [2]Department of Statistics, Texas A&M University, College Station, USA. Correspondence to: Ping-Chun Hsieh <pingchun.hsieh@tamu.edu>, Xi Liu <xiliu.tamu@gmail.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

[1]For simplicity, in this paper, we use the terms *participant*, *user*, *customer*, and *patients* interchangeably.

the outcome of the suggestion. For instance, patients will let doctors know their expectations to the treatment in physician visits. Customers are willing to inform fund managers how much money they can afford to lose. Cloud users share with the service providers their requirements of throughput performance. We consider that the outcome of following the suggestion is a random variable drawn from an unknown distribution that may vary under different contexts. If the outcome falls below the satisfaction level, the customer quits all future interactions; otherwise, the customer stays. That being said, the reneging risk is the chance that the outcome drawn from an unknown distribution falls below some threshold. Thus, learning the unknown outcome distribution plays a critical role in optimal decision making.

Learning the outcome distribution of following the suggestion can be highly challenging due to "heteroscedasticity", which means the variability of the outcome varies across the range of predictors. Many previous studies of the aforementioned applications have pointed out that the distribution of the outcome can be heteroscedastic. In treatment to a patient, it has been found the physical condition after treatment can be highly heteroscedastic (Towse et al., 2015; Buzaianu & Chen, 2018). Similarly, in portfolio selection (Omari et al., 2018; Ledoit & Wolf, 2003; Jin & Lehnert, 2018), it is even more common that the return of investing a selected portfolio is heteroscedastic. In cloud service, it has been repeatedly observed that the throughput and responses of the server can be highly heteroscedastic (Somu et al., 2018; Niu et al., 2011; Cheng & Kleijnen, 1999). In bandits setting, it means that both the mean value and the variance of the outcome depend on "context" which represents the decision and the customer. Since the reneging risk is the chance that the outcome is below the satisfaction level, accurately estimating it now requires estimation of both mean and variance. Such property makes it more difficult to learn the distribution.

While MAB and contextual bandits have been successfully applied to many sequential decision problems, they are not directly applicable to the lifetime maximization problem due to two major limitations. First, most of them neglect the phenomenon of reneging that is common in real-world applications. As a result, their objective is to maximize the accumulated rewards collected from endless interactions. As a comparison, our goal is to maximize the total number of interactions where each time of interaction faces some reneging risk. Due to that reason, conventional approaches such as LinUCB (Chu et al., 2011) will have poor performance in solving the lifetime maximization problem (see Section 5 for a comparison). Second, previous studies have usually assumed that the underlying distribution involved in the problem is homoscedastic, i.e., its variance is independent of contexts. Unfortunately, this assumption can be easily invalid due to the presence of heteroscedasticity in the motivated examples considered above, e.g., patients' health condition, portfolio return, and throughput rate.

The line of MAB research that is most relevant to the problem is bandits models with risk management, e.g., variance minimization (Sani et al., 2012) and value-at-risk maximization (Szorenyi et al., 2015; Cassel et al., 2018; Chaudhuri & Kalyanakrishnan, 2018). However, the risks those studies handle models the large fluctuation of collected rewards and have no impact on the lifetimes of bandits. This makes them unable to be applied to our problem. Another category of relevant research is conservative bandits (Kazerouni et al., 2017; Wu et al., 2016), in which a choice will only be considered if it guarantees that the *overall* performances outperforms $1-\alpha$ of baselines'. Unfortunately, our problem has a higher degree of granularity, i.e., to avoid reneging, *individual* performance (performance of each choice) is above some satisfaction level. Moreover, none of them considers data heteroscedasticity. (A more careful review and comparison are given in Section 2)

To overcome all these limitations, we propose a novel model of contextual bandits that addresses the challenges arising from reneging risk and heteroscedasticity in the lifetime maximization problem. We call the model "heteroscedastic linear bandits with reneging".

Contributions. Our research contributions are as follows:

1. Lifetime maximization is an important problem in many real-world applications but not taken into account in the existing bandit models. We investigate the two characters of the problem in aforementioned applications: reneging risk and willingness to reveal satisfaction level and propose a behavior model of reneging.

2. In view of the two characters, we formulate a novel bandits model for lifetime maximization under heteroscedasticity. It is based on our model of reneging behavior and is dubbed "heteroscedastic linear bandits with reneging."

3. To solve the proposed model, we develop a UCB-type policy, called HR-UCB, that is proved to achieve a $\mathcal{O}\big(\sqrt{T(\log(T))^3}\big)$ regret bound with high probability. We evaluate the HR-UCB via comprehensive simulations. The simulation results demonstrate that our model satisfies our expectation of regret and outperforms conventional UCB that ignores reneging and more complex model such as Episodic Reinforcement Learning (ERL).

2. Related Work

There are mainly two lines of research related to our work: bandits with risk management and conservative bandits.

Bandits with Risk Management. Reneging can be viewed as a type of risk that the decision maker tries to avoid. The risk management in bandit problems has been studied in terms of variance and quantiles. In (Sani et al., 2012), mean-variance models to handle risk are studied, where the risk

refers to the variability of collected rewards. The difference from conventional bandits is that the objective to be maximized is a linear combination of mean reward and variance. Subsequent studies (Szorenyi et al., 2015; Cassel et al., 2018) propose a quantile (value at risk) to replace the mean-variance objective. While these studies investigate optimal policies under risk, the risks they handle are different from ours, in the sense that the risks usually relate to variability of rewards and have no impact on the lifetime of bandits. Moreover, their approaches to handle the risk are based on more straightforward statistics, while, in our problem, the reneging risk is relatively complex, i.e., it comes from the probability that the outcome of following a suggestion is below a satisfaction level. Therefore, their models cannot be used to solve our problem.

Conservative Bandits. In contrast to those works, *conservative bandits* (Kazerouni et al., 2017; Wu et al., 2016) control the risk by requiring that the accumulated rewards while learning the optimal policy be above those of baselines. Similarly, in (Sun et al., 2017), each arm is associated with some risk; safety is guaranteed by requiring the accumulated risk to be below a given budget. Unfortunately, our problem has a higher degree of granularity. The participants in our problem are more sensitive to bad suggestions. One time of bad decision may incur reneging and brings the interactions to an end, e.g., one bad treatment makes a patient die. Moreover, their models assume homoscedasticity, while we allow the variance to depend on the context.

The satisfaction level in our model has the flavor of thresholding bandits. Different from us, the thresholds in the existing literature are mostly used to model reward generation. For instance, in (Abernethy et al., 2016), an action induces a unit payoff if the sampled outcome exceeds a threshold. In (Jain & Jamieson, 2018), no rewards can be collected until the total number of successes exceeds the threshold.

In terms of the problem in this paper, the most relevant one that has been studied is in (Schmit & Johari, 2018). Compared to it, our paper has three salient differences. First, it has a very different setting of reneging modeling: each decision is represented by a real number; reneging happens as long as the pulled arm exceeds a threshold. As a comparison, we represent each decision by a high-dimensional context vector; reneging happens if the outcome of following a suggestion is not satisfying. Second, it couples the reneging with the reward generation. The "rewards" in our modeling can be regarded as the lifetime while the reneging is separately captured by the outcome distribution. Third, it fails to take into account the data heteroscedasticity in the aforementioned applications. By contrast, we investigate the impacts of that and our model well addresses it.

In terms of bandits under heteroscedasticity, to the best of our knowledge, only one very recent paper discusses that (Kirschner & Krause, 2018). Compared to it, our paper has two salient differences. First, we address heteroscedasticity under the presence of reneging. The presence of reneging makes the learning problem more challenging as the learner has to always be prepared that plans for the future may not be carried out. Second, the solution in (Kirschner & Krause, 2018) is based on information directed sampling. In contrast, in this paper, we present a heteroscedastic UCB policy that is efficient, easier to implement, and can achieve sub-linear regret. The reneging problem can also be approximated by an infinite-horizon ERL problem (Modi et al., 2018; Hallak et al., 2015). Compared to it, our solution has two distinct features: (a) the reneging behavior and heteroscedasticity are explicitly addressed in our model, (b) the context information is leveraged in learning policy design.

3. Problem Formulation

In this section, we describe the formulation of the heteroscedastic linear bandits with reneging. To incorporate reneging behavior into the bandit model, we address the problem in the following stylized manner: The users arrive at the decision maker one after another and are indexed by $t = 1, 2, \cdots$. For each user t, the decision maker interacts with the user in discrete *rounds* by selecting one action in each round sequentially until the user t reneges on interacting with the decision maker. Let s_t denote the total number of rounds experienced by the user t. Note that s_t is a stopping time which depends on the reneging mechanism that will be described shortly. Since the decision maker interacts with one user at a time, all the actions and the corresponding outcomes regarding user t are determined and observed, before the next user $t+1$ arrives.

Let $\mathcal{A}$ be the set of available actions of the decision maker. Upon the arrival of each user t, the decision maker observes a set of *contexts* $\mathcal{X}_t = \{x_{t,a}\}_{a\in\mathcal{A}}$, where each context $x_{t,a} \in \mathcal{X}_t$ summarizes the pair-wise relationship[2] between the user t and the action a. Without loss of generality, we assume that for any user t and any action a, we have $\|x_{t,a}\|_2 \leq 1$, where $\|\cdot\|_2$ denotes the ℓ_2-norm. After observing the contexts, the decision maker selects an action $a \in \mathcal{A}$ and observes a random outcome $r_{t,a}$. We assume that the outcomes $r_{t,a}$ are conditionally independent random variables given the contexts and are drawn from an outcome distribution that satisfies:

$$r_{t,a} := \theta_*^\top x_{t,a} + \varepsilon(x_{t,a}) \tag{1}$$

$$\varepsilon(x_{t,a}) \sim \mathcal{N}\big(0, \sigma^2(x_{t,a})\big) \tag{2}$$

$$\sigma^2(x_{t,a}) := f(\phi_*^\top x_{t,a}), \tag{3}$$

[2]For example, in recommender systems, one way to construct a such pair-wise context is to concatenate the feature vectors of each individual user and each individual action.

where $\mathcal{N}(0, \sigma^2)$ denotes the Gaussian distribution with zero mean and variance σ^2, and $\theta_*, \phi_* \in \mathbb{R}^d$ are unknown, but known to have the norm bounds as $||\theta_*||_2 \leq 1$ and $||\phi_*||_2 \leq L$. Although, for simplicity of discussion, here we focus on Gaussian noise, all of our analysis can be extended to sub-Gaussian outcome distribution of the form $\psi_\sigma(x) = (1/\sigma)\psi((x-\mu)/\sigma)$, where ψ is a known sub-Gaussian density with unknown parameters μ, σ. This family includes truncated distributions and mixtures, thus allowing multi-modality and skewness. The parameter vectors $\theta_* \in \mathbb{R}^d$ and $\phi_* \in \mathbb{R}^d$ will be learned by the decision maker during interactions with the users. The function $f(\cdot) : \mathbb{R} \to \mathbb{R}$ is assumed to be a known linear function with a finite positive slope M_f such that $f(z) \geq 0$, for all $z \in [-L, L]$. One example that satisfies the above conditions is $f(z) = z + L$. Note that the mean and variance of the outcome distribution satisfy

$$\mathbb{E}[r_{t,a}|x_{t,a}] := \theta_*^\top x_{t,a}, \tag{4}$$

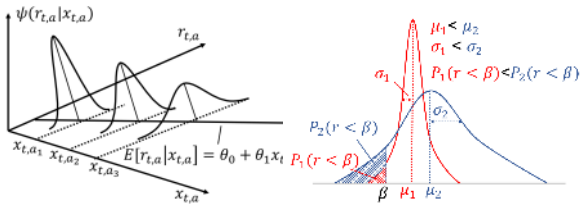
$$\mathbb{V}[r_{t,a}|x_{t,a}] := f(\phi_*^\top x_{t,a}). \tag{5}$$

Since $\phi_*^\top x_{t,a}$ is bounded over all possible ϕ_* and $x_{t,a}$, we know that $f(\phi_*^\top x_{t,a})$ is also bounded, i.e. $f(\phi_*^\top x_{t,a}) \in [\sigma_{\min}^2, \sigma_{\max}^2]$ for some $\sigma_{\min}, \sigma_{\max} > 0$, for all ϕ_* and $x_{t,a}$ defined above. This also implies that $\varepsilon(x_{t,a})$ is $\sigma_{\max}^2$-sub-Gaussian, for all $x_{t,a}$.

The minimal expectation in an interaction of a user is characterized by its *satisfaction level*. Let $\beta_t \in \mathbb{R}$ denote the satisfaction level of user t. We assume that satisfaction levels of users, like the pair-wise contexts, are available before interacting with them. Denote by $r_t^{(i)}$ the observed outcome at round i of user t. When $r_t^{(i)}$ is below β_t, reneging occurs and the user drops out from any future interaction. Suppose that at round i, action a is selected for user t, then the risk that reneging occurs is

$$\mathbb{P}(r_t^{(i)} < \beta_t | x_{t,a}) = \Phi\Big(\frac{\beta_t - \theta_*^\top x_{t,a}}{\sqrt{f(\phi_*^\top x_{t,a})}}\Big), \tag{6}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) for $\mathcal{N}(0,1)$. Without loss of generality, we also assume that β_t is lower bounded by $-B$ for some $B > 0$. Recall that s_t denotes the number of rounds experienced by user t. Given the reneging behavior modeled above, s_t is the stopping time that represents the first time that the outcome $r_t^{(i)}$ is below the satisfaction level β_t, i.e. $s_t := \min\{i : r_t^{(i)} < \beta_t\}$. Illustrative examples of heteroscedasticity and reneging risk are shown in Figure 1. In Figure 1(a), the variance of the outcome distribution gradually increases as the value of the one-dimensional context $x_{t,a}$ increases. Figure 1(b) shows the outcome distributions of the two actions for a user. Specifically, the outcome distribution P_1 has mean μ_1 and variance σ_1^2, and mean μ_2 and variance σ_2^2 for P_2. As the two distributions correspond to the same user (but for different actions), they face the same satisfaction level β. In this example, the reneging risk $P_2(r < \beta)$ (the blue shaded area) is higher than $P_1(r < \beta)$ (the red shaded area).



(a) Example of heteroscedasticity (b) Example of reneging risk

Figure 1. Illustrated examples for heteroscedasticity and reneging risk under presence of heteroscedasticity. ($\psi(\cdot)$ is the probability density function.)

A policy $\pi \in \Pi$ is a rule for selecting an action at each round for a user based on the preceding interactions with that user and other users, where Π denotes the set of all admissible policies. Let $\pi_t = \{x_{t,1}, x_{t,2}, \cdots\}$ denote the sequence of contexts that correspond to the actions for user t under policy π. Let $\overline{R}_t^\pi$ denote the expected lifetime of user t under the action sequence π_t. Then the total expected lifetime of T users can be represented by $\mathcal{R}^\pi(T) = \sum_{t=1}^T \overline{R}_t^{\pi_t}$. Define π^* as the optimal policy in terms of total expected lifetime among admissible policies, i.e. $\pi^* = \arg\max_{\pi \in \Pi} \mathcal{R}^\pi(T)$. We are ready to define the *pseudo regret* of the heteroscedastic linear bandits with reneging for a policy π as

$$\text{Regret}_T := \mathcal{R}^{\pi^*}(T) - \mathcal{R}^\pi(T). \tag{7}$$

The objective of the decision maker is to learn a policy that achieves as minimal a regret as possible.

4. Algorithms and Results

In this section, we present a UCB-type algorithm for heteroscedastic linear bandits with reneging. We start by introducing general results on heteroscedastic regression.

4.1. Heteroscedastic Regression

In this section, we consider a general regression problem with heteroscedasticity.

4.1.1. Generalized Least Squares Estimators

With a slight abuse of notation, let $\{(x_i, r_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n$ be a sequence of n pairs of context and outcome that are realized by a user's actions. Recall from (1)-(3) that $r_i = \theta_*^\top x_i + \varepsilon(x_i)$ and $\varepsilon(x_i) \sim \mathcal{N}\big(0, f(\phi_*^\top x_i)\big)$ with unknown parameters θ_* and ϕ_*. Note that, given the contexts $\{x_i\}_{i=1}^n$, $\varepsilon(x_1), \cdots, \varepsilon(x_n)$ are mutually independent. Let $r = (r_1, \cdots, r_n)^\top$ and $\varepsilon = (\varepsilon(x_1), \cdots, \varepsilon(x_n))$ be the row vectors of the n outcome realizations and the deviations from the mean, respectively. Let $\boldsymbol{X}_n$ be an $n \times d$ matrix in which the i-th row is $x_i^\top$, for all $1 \leq i \leq n$.

We use $\widehat{\theta}_n, \widehat{\phi}_n \in \mathbb{R}^d$ to denote the estimators of θ_* and ϕ_* based on the observations $\{(x_i, r_i)\}_{i=1}^n$, respectively. Moreover, define the estimated residual with respect to $\widehat{\theta}_n$ as $\widehat{\varepsilon}(x_i) = r_i - \widehat{\theta}_n^\top x_i$. Let $\widehat{\varepsilon} = (\widehat{\varepsilon}(x_1), \cdots, \widehat{\varepsilon}(x_n))^\top$. Let $\boldsymbol{I}_d$ denote the $d \times d$ identity matrix, and let $z_1 \circ z_2$ denote the Hadamard product of any two vectors z_1, z_2. We consider the *generalized least squares estimators* (GLSE) (Wooldridge, 2015) as

$$\widehat{\theta}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}_n^\top r, \tag{8}$$

$$\widehat{\phi}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\right)^{-1} \boldsymbol{X}_n^\top f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}), \tag{9}$$

where $\lambda > 0$ is some regularization parameter and $f^{-1}(\widehat{\varepsilon} \circ \widehat{\varepsilon}) = (f^{-1}(\widehat{\varepsilon}(x_1)^2), \cdots, f^{-1}(\widehat{\varepsilon}(x_n)^2))^\top$ is the pre-image of the vector $\widehat{\varepsilon} \circ \widehat{\varepsilon}$.

Remark 1 Note that in (8), $\widehat{\theta}_n$ is the conventional ridge regression estimator. On the other hand, to obtain an estimator $\widehat{\phi}_n$, (9) still follows the ridge regression approach, but with two additional steps: (i) derive the estimated residual $\widehat{\varepsilon}$ based on $\widehat{\theta}_n$, and (ii) apply the map $f^{-1}(\cdot)$ on the square of $\widehat{\varepsilon}$. Conventionally, GLSE is utilized to improve the efficiency of estimating θ_* under heteroscedasticity (e.g. Chapter 8.2 of (Wooldridge, 2015)). In our problem, we use GLSE to jointly learn θ_* and ϕ_* and thereby establish regret guarantees. However, it is not immediately clear how to obtain finite-time results regarding the confidence set for $\widehat{\phi}_n$. This issue will be addressed in Section 4.1.2.

4.1.2. Confidence Sets for GLSE

In this section, we discuss the confidence sets for the estimators $\widehat{\theta}_n$ and $\widehat{\phi}_n$ described above. To simplify notation, we define a $d \times d$ matrix $\boldsymbol{V}_n$ as

$$\boldsymbol{V}_n = \left(\boldsymbol{X}_n^\top \boldsymbol{X}_n + \lambda \boldsymbol{I}_d\right). \tag{10}$$

A confidence set for $\widehat{\theta}_t$ was introduced in (Abbasi-Yadkori et al., 2011). For convenience, we restate these elegant results in the following lemma.

Lemma 1 (Theorem 2 in (Abbasi-Yadkori et al., 2011)) *For all $n \in \mathbb{N}$, define*

$$\alpha_n^{(1)}(\delta) = \sigma_{\max}^2 \sqrt{d \log\left(\frac{n+\lambda}{\delta\lambda}\right)} + \lambda^{1/2}. \tag{11}$$

For any $\delta > 0$, we have

$$\mathbb{P}\left\{\left\|\widehat{\theta}_n - \theta_*\right\|_{\boldsymbol{V}_n} \le \alpha_n^{(1)}(\delta), \forall n \in \mathbb{N}\right\} \ge 1 - \delta, \tag{12}$$

where $\|x\|_{\boldsymbol{V}_n} = \sqrt{x^\top \boldsymbol{V}_n x}$ is the induced vector norm of vector x with respect to $\boldsymbol{V}_n$.

Next, we derive the confidence set for $\widehat{\phi}_n$. Define

$$\alpha^{(2)}(\delta) = \sqrt{2d(\sigma_{\max}^2)^2\left(\left(\frac{1}{C_2}\ln(\frac{C_1}{\delta})\right)^2 + 1\right)}, \tag{13}$$

$$\alpha^{(3)}(\delta) = \sqrt{2d\sigma_{\max}^2 \ln\left(\frac{d}{\delta}\right)}, \tag{14}$$

where C_1 and C_2 are some universal constants that will be described in Lemma 3. The following is the main theorem on the confidence set for $\widehat{\phi}_n$.

Theorem 1 *For all $n \in \mathbb{N}$, define*

$$\rho_n(\delta) = \frac{1}{M_f}\Big\{\alpha_n^{(1)}(\tfrac{\delta}{3})\big(\alpha_n^{(1)}(\tfrac{\delta}{3}) + 2\alpha^{(3)}(\tfrac{\delta}{3})\big) \tag{15}$$

$$+ \alpha^{(2)}(\tfrac{\delta}{3})\Big\} + L^2\lambda^{1/2}. \tag{16}$$

For any $\delta > 0$, with probability at least $1 - 2\delta$, we have

$$\left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n} \le \rho_n(\frac{\delta}{n^2}) = \mathcal{O}\Big(\log(\frac{1}{\delta}) + \log n\Big), \forall n \in \mathbb{N}. \tag{17}$$

Remark 2 As the estimator $\hat{\phi}_n$ depends on the residual term $\hat{\varepsilon}$, which involves the estimator $\hat{\theta}_n$, it is expected that the convergence speed of $\hat{\phi}_n$ would be no larger than that of $\hat{\theta}_n$. Based on Theorem 1 along with Lemma 1, we know that under GLSE, $\hat{\phi}_n$ converges to the true value at a slightly slower rate than $\hat{\theta}_n$.

To demonstrate the main idea behind Theorem 1, we highlight the proof in the following Lemma 2-5. We start by taking the inner products of an arbitrary vector x with $\widehat{\phi}_n$ and ϕ_* to quantify the difference between $\widehat{\phi}_t$ and ϕ_*.

Lemma 2 *For any $x \in \mathbb{R}^d$, we have*

$$|x^\top \widehat{\phi}_n - x^\top \widehat{\phi}_*| \le \|x\|_{\boldsymbol{V}_n^{-1}} \left\{\lambda \|\phi_*\|_{\boldsymbol{V}_n^{-1}}\right. \tag{18}$$

$$+ \left\|\boldsymbol{X}_n^\top\left(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n\phi_*\right)\right\|_{\boldsymbol{V}_n^{-1}} \tag{19}$$

$$+ \frac{2}{M_f}\left\|\boldsymbol{X}_n^\top\left(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\right)\right\|_{\boldsymbol{V}_n^{-1}} \tag{20}$$

$$+ \frac{1}{M_f}\left\|\boldsymbol{X}_n^\top\left(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\right)\right\|_{\boldsymbol{V}_n^{-1}}\Big\}. \tag{21}$$

Proof. The proof is provided in Appendix A.1. □

Based on Lemma 2, we provide upper bounds for the three terms in (19)-(21) separately as follows.

Lemma 3 *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$M_f \left\|\boldsymbol{X}_n^\top\left(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n\phi_*\right)\right\|_{\boldsymbol{V}_n^{-1}} \le \alpha^{(2)}(\delta). \tag{22}$$

Proof. We highlight the main idea of the proof. Recall that $\varepsilon(x_i) \sim \mathcal{N}(0, \phi_*^\top x_i)$. Therefore, $\varepsilon(x_i)^2$ is a χ_1^2-distribution with a scaling of $f(\phi_*^\top x_i)$. Hence, each element

in $(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n\phi_*)$ has zero mean. Moreover, we observe that $\left\|\boldsymbol{X}_n^\top\left(f^{-1}(\varepsilon \circ \varepsilon) - \boldsymbol{X}_n\phi_*\right)\right\|_{\boldsymbol{V}_n{}^{-1}}$ is quadratic. Since the χ_1^2-distribution is sub-exponential, we utilize a proper tail inequality for quadratic forms of sub-exponential distributions to derive an upper bound. The complete proof is provided in Appendix A.2. □

Next, we derive an upper bound for (20).

Lemma 4 *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\left\|\boldsymbol{X}_n^\top\left(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\right)\right\|_{\boldsymbol{V}_n{}^{-1}} \le \alpha_n^{(1)}(\delta) \cdot \alpha^{(3)}(\delta). \quad (23)$$

Proof. The main challenge is that (23) involves the product of the residual ε and the estimation error $\theta_* - \widehat{\theta}_n$. Through some manipulation, we can decouple ε from $\left\|\boldsymbol{X}_n^\top\left(\varepsilon \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\right)\right\|_{\boldsymbol{V}_n{}^{-1}}$ and apply a proper tail inequality for quadratic forms of sub-Gaussian distributions. The complete proof is provided in Appendix A.3. □

Next, we provide an upper bound for (21).

Lemma 5 *For any $n \in \mathbb{N}$, for any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\left\|\boldsymbol{X}_n^\top\left(\boldsymbol{X}_n(\theta_* - \widehat{\theta}_n) \circ \boldsymbol{X}_n(\theta_* - \widehat{\theta}_n)\right)\right\|_{\boldsymbol{V}_n{}^{-1}} \le (\alpha_n^{(1)}(\delta))^2. \quad (24)$$

Proof. Since (24) does not involve ε, we can simply reuse the results in Lemma 1 through some manipulation of (24). The complete proof is provided in Appendix A.4. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1. We use $\lambda_{\min}(\cdot)$ to denote the smallest eigenvalue of a square symmetric matrix. Recall that $\boldsymbol{V}_n = \lambda\boldsymbol{I}_d + \boldsymbol{X}_n^\top\boldsymbol{X}_n$ is positive definite for all $\lambda > 0$. We have

$$\|\phi_*\|^2_{\boldsymbol{V}_n{}^{-1}} \le \|\phi_*\|_2^2/\lambda_{\min}(\boldsymbol{V}_n) \le \|\phi_*\|_2^2/\lambda \le L^2/\lambda. \quad (25)$$

By (25) and Lemmas 2-5, we know that for a given n and a given $\delta_n > 0$, with probability at least $1-\delta_n$, we have

$$|x^\top\widehat{\phi}_n - x^\top\widehat{\phi}_*| \le \|x\|_{\boldsymbol{V}_n{}^{-1}} \cdot \rho_n(\delta_n). \quad (26)$$

Note that (26) holds for any $x \in \mathbb{R}^d$. By substituting $x = \boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)$ into (26), we have

$$\left\|\widehat{\phi}_n - \phi_*\right\|^2_{\boldsymbol{V}_n} \le \left\|\boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)\right\|_{\boldsymbol{V}_n{}^{-1}} \cdot \rho_n(\delta_n). \quad (27)$$

Since $\left\|\boldsymbol{V}_n(\widehat{\phi}_n - \phi_*)\right\|_{\boldsymbol{V}_n{}^{-1}} = \left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n}$, we know for a given n and $\delta_n > 0$, with probability at least $1-\delta_n$,

$$\left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n} \le \rho_n(\delta_n). \quad (28)$$

Finally, to obtain a uniform bound, we simply choose $\delta_n = \delta/(n^2)$ and apply the union bound to (28) over all $n \in \mathbb{N}$. Note that $\sum_{n=1}^\infty \delta_n = \sum_{n=1}^\infty \delta/n^2 = \frac{\pi^2}{6}\delta < 2\delta$. Therefore, with probability at least $1-2\delta$, for all $n \in \mathbb{N}$, $\left\|\widehat{\phi}_n - \phi_*\right\|_{\boldsymbol{V}_n} \le \rho_n\left(\frac{\delta}{n^2}\right)$. □

4.2. Heteroscedastic UCB Policy

In this section, we formally introduce the proposed policy based on the heteroscedastic regression in Section 4.1.

4.2.1. An Oracle Policy

In this section, we consider a policy which has access to an oracle with full knowledge of θ_* and ϕ_*. Consider T users that arrive sequentially. Let $\pi_t^{\text{oracle}} = \{x_{t,1}^*, x_{t,2}^*, \cdots\}$ be the sequence of contexts that correspond to the actions for the user t under an oracle policy π^{oracle}. The oracle policy $\pi^{\text{oracle}} = \{\pi_t^{\text{oracle}}\}$ is constructed by choosing

$$\pi_t^{\text{oracle}} = \arg\max_{\widetilde{x}_t = \{\widetilde{x}_{t,1}, \widetilde{x}_{t,2}\cdots\}} R_t^{\widetilde{x}_t}, \quad (29)$$

for each t. Due to the construction in (29), we know that π^{oracle} achieves the largest possible expected lifetime for each user t, and is hence optimal in terms of pseudo-regret defined in Section 3. By using an one-step optimality argument, it is easy to verify that π^{oracle} is a fixed policy for each user t, i.e. $x_{t,i} = x_{t,j}$, for all $i, j \ge 1$. Let $\overline{R}_t^*$ denote the expected lifetime of user t under π^{oracle}. We have

$$\overline{R}_t^* = \left(\Phi\left(\frac{\beta_t - \theta_*^\top x_t^*}{\sqrt{f(\phi_*^\top x_t^*)}}\right)\right)^{-1}. \quad (30)$$

Next, we derive a useful property regarding (30). For any given $\beta \in [-B, \infty)$, define the function $h_\beta : [-1,1] \times [\sigma_{\min}^2, \sigma_{\max}^2] \to \mathbb{R}$ as

$$h_\beta(u, v) = \left(\Phi\left(\frac{\beta - u}{\sqrt{f(v)}}\right)\right)^{-1}. \quad (31)$$

Note that for any given $x \in \mathcal{X}$, $h_\beta({\theta_*}^\top x, {\phi_*}^\top x)$ equals the expected lifetime of a single user with threshold β if a fixed action with context x is chosen under parameters θ_*, ϕ_*. We show that $h_\beta(\cdot,\cdot)$ has the following nice property.

Theorem 2 *Let $\boldsymbol{M}$ be a $d \times d$ invertible matrix. For any $\theta_1, \theta_2 \in \mathbb{R}^d$ with $\|\theta_1\| \le 1$, $\|\theta_2\| \le 1$, for any $\phi_1, \phi_2 \in \mathbb{R}^d$ with $\|\phi_1\| \le L$, $\|\phi_2\| \le L$, for any $\beta \in [-B, \infty)$, $\forall x \in \mathcal{X}$,*

$$h_\beta\left(\theta_2^\top x, \phi_2^\top x\right) - h_\beta\left(\theta_1^\top x, \phi_1^\top x\right) \le \quad (32)$$

$$\left(C_3\left\|\theta_2 - \theta_1\right\|_{\boldsymbol{M}} + C_4\left\|\phi_2 - \phi_1\right\|_{\boldsymbol{M}}\right) \cdot \|x\|_{\boldsymbol{M}^{-1}}, \quad (33)$$

where C_3 and C_4 are some finite positive constants that are independent of $\theta_1, \theta_2, \phi_1, \phi_2$, and β.

Proof. The main idea is to apply first-order approximation under Lipschitz continuity of $h_\beta(\cdot,\cdot)$. The detailed proof is provided in Appendix A.5. □

4.2.2. The HR-UCB Policy

To begin with, we introduce an upper confidence bound based on the GLSE described in Section 4.1. Note that the

results in Theorem 1 depend on the size of the set of context-outcome pairs. Moreover, in our bandit model, the number of rounds of each user is a stopping time and can be arbitrarily large. To address this, we propose to actively maintain a *regression sample set* $\mathcal{S}$ through a function $\Gamma(t)$. Specifically, we let the size of $\mathcal{S}$ grow at a proper rate regulated by $\Gamma(t)$. One example is to choose $\Gamma(t) = Kt$ for some constant $K \geq 1$. Since each user will play for at least one round, we know $|\mathcal{S}|$ is at least t after interacting with t users. We use $\mathcal{S}(t)$ to denote the regression sample set right after the departure of user t. Moreover, let $\boldsymbol{X}_t$ be the matrix in which the rows are composed by the contexts of all the elements in $\mathcal{S}(t)$. Similar to (10), we define $\boldsymbol{V}_t = \boldsymbol{X}_t^\top \boldsymbol{X}_t + \lambda \boldsymbol{I}_d$, for all $t \geq 1$. To simplify notation, we also define

$$\xi_t(\delta) := C_3 \alpha^{(1)}_{|\mathcal{S}(t)|}(\delta) + C_4 \rho_{|\mathcal{S}(t)|}(\delta/|\mathcal{S}(t)|^2). \tag{34}$$

For any $x \in \mathcal{X}$, we define the upper confidence bound as

$$Q^{\text{HR}}_{t+1}(x) = h_{\beta_{t+1}}\big(\widehat{\theta}_t^\top x, \widehat{\phi}_t^\top x\big) + \xi_t(\delta) \cdot \|x\|_{\boldsymbol{V}_t^{-1}}. \tag{35}$$

Note that the exploration over users, guaranteeing sublinear regret under heteroscedasticity, is handled by encoding the confidence bound in Q^{HR}_t so that later users with similar contexts are treated differently. Next, we show that $Q^{\text{HR}}_t(x)$ is indeed an upper confidence bound.

Algorithm 1 The HR-UCB Policy

$\mathcal{S} \leftarrow \varnothing$, action set $\mathcal{A}$, function $\Gamma(t)$, and T
for each user $t = 1, 2, \cdots, T$ **do**
  observe $x_{t,a}$ for all $a \in \mathcal{A}$ and reset $i \leftarrow 1$
  while user t stays **do**
    $\pi_t^{(i)} = \arg\max_{x_{t,a} \in \mathcal{X}_t} Q^{\text{HR}}_t(x_{t,a})$ (ties are broken arbitrarily)
    apply the action $\pi_t^{(i)}$ and observe the outcome $r_t^{(i)}$ and if the reneging event occurs
    if $|\mathcal{S}| < \Gamma(t)$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup \{(x_{t,\pi_t^{(i)}}, r_t^{(i)})\}$
    end if
    $i \leftarrow i + 1$
  end while
  update $\widehat{\theta}_t$ and $\widehat{\phi}_t$ by (8)-(9) based on $\mathcal{S}$
end for

Lemma 6 *If the confidence set conditions (12) and (17) are satisfied, then for any $x \in \mathcal{X}$,*

$$0 \leq Q^{\text{HR}}_{t+1}(x) - h_{\beta_{t+1}}\big(\theta_*^\top x, \phi_*^\top x\big) \leq 2\xi_t(\delta)\, \|x\|_{\boldsymbol{V}_t^{-1}}.$$

Proof. The proof is provided in Appendix A.6. □

Now, we formally introduce the HR-UCB algorithm.

- For each user t, HR-UCB observes the contexts of all available actions and then chooses an action based on the indices Q^{HR}_t that depend on $\widehat{\theta}_t$ and $\widehat{\phi}_t$. To derive these estimators by (8) and (9), HR-UCB actively maintains a sample set $\mathcal{S}$, whose size is regulated by a function $\Gamma(t)$.
- After applying an action, HR-UCB observes the corresponding outcome and the reneging event if any. The current context-outcome pair will be added to $\mathcal{S}$ only if the size of $\mathcal{S}$ is less than $\Gamma(t)$.
- Based on the regression sample set $\mathcal{S}$, HR-UCB updates $\widehat{\theta}_t$ and $\widehat{\phi}_t$ right after the departure of each user.

The complete algorithm is shown in Algorithm 1.

Remark 3 In Algorithm 1, $\widehat{\theta}_t$ and $\widehat{\phi}_t$ are updated right after the departure of each user. Alternatively, $\widehat{\theta}_t$ and $\widehat{\phi}_t$ can be updated whenever $\mathcal{S}$ is updated. While this alternative makes slightly better use of the observations, it also incurs more computation overhead. For ease of exposition, we focus on the "lazy-update" version presented in Algorithm 1.

4.3. Regret Analysis

In this section, we provide regret analysis for HR-UCB.

Theorem 3 *Under HR-UCB, with probability at least $1-\delta$, the pseudo regret is upper bounded as*

$$\text{Regret}_T \leq \sqrt{8\xi_T^2(\frac{\delta}{3})T \cdot d \log\Big(\frac{T+\lambda d}{\lambda d}\Big)} \tag{36}$$

$$= \mathcal{O}\bigg(\sqrt{T \log \Gamma(T) \cdot \Big(\log\big(\Gamma(T)\big) + \log(\frac{1}{\delta})\Big)^2}\bigg). \tag{37}$$

By choosing $\Gamma(T) = KT$ with a constant $K > 0$, we have

$$\text{Regret}_T = \mathcal{O}\bigg(\sqrt{T \log T \cdot \Big(\log T + \log(\frac{1}{\delta})\Big)^2}\bigg). \tag{38}$$

Proof. The proof is provided in Appendix A.7. □

Theorem 3 presents a high-probability regret bound. To derive an expected regret bound, we can set $\delta = 1/T$ in (37) and get $\mathcal{O}(\sqrt{T(\log T)^3})$. Also note that the upper bound (36) depends on $\sigma_{\max}$ only through the pre-constant of ξ_T.

Remark 4 A policy that always assumes $\sigma_{\max}$ as variance tends to choose the action with the largest mean reward since it implies a smaller reneging probability. As a result, such type of policy incurs linear regret. This will be further demonstrated via simulations in Section 5.

Remark 5 The regret proof still goes through for sub-Gaussian noise by (a) reusing the same sub-exponential concentration inequality in Lemma A.1 since the square of an sub-Gaussian distribution is sub-exponential, (b) replacing the Gaussian concentration inequality in Lemma A.3 with a sub-Gaussian one, and (c) deriving ranges of the first two derivatives of sub-Gaussian CDF.

Remark 6 The assumption that β_t is known can be relaxed to the case where only the distribution of β_t is known. The analysis can be adapted to this case by (a) rewriting the reneging probability in (6) and $h_\beta(u, v)$ in (31) via integration over distribution of β_t, (b) deriving the corresponding expected lifetime under oracle policy in (30), and (c) reusing

Theorem 1 and Lemma 1 as the GLSE does not rely on the knowledge of β_t.

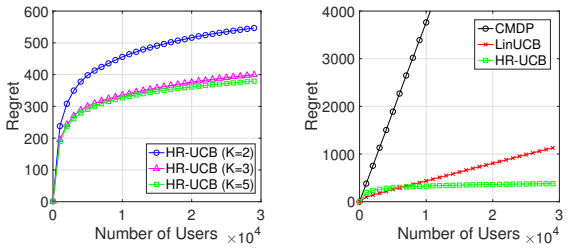
Remark 7 We briefly discuss the difference between our regret bound and the regret bounds of other related settings. Note that if the satisfaction level $\beta_t = \infty$ for all t, then all the users will quit after exactly one round. This corresponds to the conventional contextual bandits setting (e.g. homoscedastic case (Chu et al., 2011) and heteroscedastic case (Kirschner & Krause, 2018)). In this degenerate case, our regret bound is $\mathcal{O}(\sqrt{T(\log T)} \cdot \log T)$, which has an additional factor $\log T$ resulting from the heteroscedasticity.

5. Simulation Results

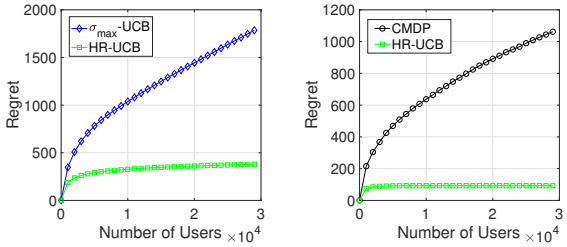
In this section, we evaluate the performance of HR-UCB. We consider 20 actions available to the decision maker. For simplicity, the context of each user-action pair is designed to be a four-dimensional vector, which is drawn uniformly at random from a unit ball. For the mean and variance of the outcome distribution, we set $\theta_* = [0.6, 0.5, 0.5, 0.3]^\top$ and $\phi_* = [0.5, 0.2, 0.8, 0.9]^\top$, respectively. We consider the function $f(x) = x + L$ with $L = 2$ and $M_f = 1$. The acceptance level of each user is drawn uniformly at random from the interval $[-1, 1]$. We set $T = 30000$ throughout the simulations. For HR-UCB, we set $\delta = 0.1$ and $\lambda = 1$. All the results in this section are the average of 20 simulation trials. Recall that K denotes the growth rate of the regression sample set for HR-UCB. We start by evaluating the pseudo regrets of HR-UCB under different K, as shown in Figure 2a. Note that HR-UCB achieves a sublinear regret regardless of K. The effect of K is only reflected when the number of users is small. Specifically, a smaller K induces a slightly higher regret since it requires more users in order to accurately learn the parameters. Based on Figure 2a, we set $K = 5$ for the rest of the simulations.

Next, we compare the HR-UCB policy with the well-known LinUCB policy (Li et al., 2010) and the CMDP policy (Modi et al., 2018). Figure 2b shows the pseudo regrets under LinUCB, CMDP and HR-UCB. LinUCB achieves a linear regret because it does not take into account the heteroscedasticity of the outcome distribution in the existence of reneging. For each user, LinUCB simply chooses the action with the largest predicted mean of the outcome distribution. The regret attained by CMDP policy also appears linear. This is because CMDP handles contexts by partitioning the context space and then learning each partition-induced MDP separately. Due to the continuous context space, the CMDP policy requires numerous partitions as well as plentiful exploration for all MDPs. To make the comparison more fair, we consider a simpler setting with a discrete context space of size 10 and only 2 actions (with other parameters unchanged). In this setting, Figure 2d shows that the regret attained by CMDP is still much larger than that by HR-UCB and thereby shows the advantage of the proposed solution.

As mentioned in Section 4.3, a policy (denoted by $\sigma_{\max}$-UCB) that always assumes $\sigma_{\max}$ as variance tends to choose the action with the largest mean and thus incurs linear regret. We demonstrate the statement in experiments shown by Figure 2c, where the $\sigma_{\max}$-UCB policy attains a linear regret vs. HR-UCB achieves a sublinear and much smaller regret. Through simulations, we validate that HR-UCB achieves the regret performance as discussed in Section 4.



(a) Pseudo regrets: HR-UCB with different K.
(b) Pseudo regrets: LinUCB, CMDP and HR-UCB ($K = 5$).



(c) Pseudo regrets: $\sigma_{\max}$-UCB and HR-UCB ($K = 5$).
(d) Pseudo regrets: CMDP and HR-UCB ($K = 5$).

Figure 2. Comparison of pseudo regrets.

6. Concluding Remarks

There are several ways to extend the studies in this paper. First, the techniques used to estimate heteroscedastic variance and establishing sub-linear regret under the presence of heteroscedasticity can be extended to other variance-sensitive bandit problems, e.g., risk-averse bandits and thresholding bandits. Second, the studies can be easily adapted to another objective - maximizing total collected rewards by: (a) replacing $h_\beta(u, v)$ in (30) with $\hat{h}_\beta(u, v) = u \cdot h_\beta(u, v)$, (b) reusing Theorem 1 and Lemma 1, and (c) making minor changes to constants C_3, C_4 in (33). Third, another promising extension is to use active-learning to update sample set $\mathcal{S}$ (Riquelme et al., 2017). To provide theoretical guarantees, these active-learning approaches often assume that arriving contexts are i.i.d. In contrast, since that assumption can be easily invalid (e.g., it is adversarial), we establish the regret bound without making any such assumption. Finally, in this paper, the problem of knowledge transfer across users is given more importance than learning for a single user. This is because, compared to the population of potential users, a user's lifetime is mostly short. Therefore, another possible extension is to take into account the exploration during the lifetime of each individual user.

Acknowledgements

This material is based upon work partially supported by NSF under Science & Technology Center Grant CCF-0939370, and Texas A&M University under the Presidents Excellence Funds X Grants Program. We would like to thank all reviewers and Professor P. S. Sastry for their insightful suggestions!

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.

Abernethy, J. D., Amin, K., and Zhu, R. Threshold bandits, with and without censored feedback. In *Advances In Neural Information Processing Systems*, pp. 4889–4897, 2016.

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

Bubeck, S. et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Buzaianu, E. M. and Chen, P. A two-stage design for comparative clinical trials: The heteroscedastic solution. *Sankhya B*, 80(1):151–177, 2018.

Cassel, A., Mannor, S., and Zeevi, A. A general approach to multi-armed bandits under risk criteria. In *Annual Conference on Learning Theory*, pp. 1295–1306, 2018.

Chaudhuri, A. R. and Kalyanakrishnan, S. Quantile-regret minimisation in infinitely many-armed bandits. In *Association for Uncertainty in Artificial Intelligence*, 2018.

Cheng, R. C. H. and Kleijnen, J. P. C. Improved design of queueing simulation experiments with highly heteroscedastic responses. *Oper. Res.*, 47(5):762–777, May 1999. ISSN 0030-364X. doi: 10.1287/opre.47.5.762.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Ding, W., Qiny, T., Zhang, X.-D., and Liu, T.-Y. Multi-armed bandit with budget constraint and variable costs. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, pp. 232–238. AAAI Press, 2013.

Erdős, L., Yau, H.-T., and Yin, J. Bulk universality for generalized Wigner matrices. *Probability Theory and Related Fields*, 154(1-2):341–407, 2012.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Huo, X. and Fu, F. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

Jain, L. and Jamieson, K. Firing bandits: Optimizing crowdfunding. In *International Conference on Machine Learning*, pp. 2211–2219, 2018.

Jin, X. and Lehnert, T. Large portfolio risk management and optimal portfolio allocation with dynamic elliptical copulas. *Dependence Modeling*, 6(1):19–46, 2018.

Kazerouni, A., Ghavamzadeh, M., Abbasi, Y., and Van Roy, B. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, pp. 3910–3919, 2017.

Kirschner, J. and Krause, A. Information directed sampling and bandits with heteroscedastic noise. In *Annual Conference on Learning Theory*, pp. 358–384, 2018.

Ledoit, O. and Wolf, M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5): 603–621, 2003.

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.

Liu, X., Xie, M., Wen, X., Chen, R., Ge, Y., Duffield, N., and Wang, N. A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games. In *2018 IEEE International Conference on Data Mining*, pp. 277–286. IEEE, 2018.

McHugh, N., Baker, R. M., Mason, H., Williamson, L., van Exel, J., Deogaonkar, R., Collins, M., and Donaldson, C. Extending life for people with a terminal illness: a moral right and an expensive death? exploring societal perspectives. *BMC Medical Ethics*, 16(1), 2015.

Modi, A., Jiang, N., Singh, S., and Tewari, A. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pp. 597–618, 2018.

Niu, D., Li, B., and Zhao, S. Understanding demand volatility in large vod systems. In *Proceedings of the 21st international workshop on Network and operating systems*

support for digital audio and video, pp. 39–44. ACM, 2011.

Omari, C. O., Mwita, P. N., and Gichuhi, A. W. Currency portfolio risk measurement with generalized autoregressive conditional heteroscedastic-extreme value theory-copula model. *Journal of Mathematical Finance*, 8(02): 457, 2018.

Riquelme, C., Johari, R., and Zhang, B. Online active linear regression via thresholding. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Sani, A., Lazaric, A., and Munos, R. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pp. 3275–3283, 2012.

Schmit, S. and Johari, R. Learning with abandonment. In *International Conference on Machine Learning*, pp. 4516–4524, 2018.

Somu, N., MR, G. R., Kalpana, V., Kirthivasan, K., and VS, S. S. An improved robust heteroscedastic probabilistic neural network based trust prediction approach for cloud service selection. *Neural Networks*, 108:339–354, 2018.

Sun, W., Dey, D., and Kapoor, A. Safety-aware algorithms for adversarial contextual bandit. In *International Conference on Machine Learning*, pp. 3280–3288, 2017.

Szorenyi, B., Busa-Fekete, R., Weng, P., and Hüllermeier, E. Qualitative multi-armed bandits: A quantile-based approach. In *32nd International Conference on Machine Learning*, pp. 1660–1668, 2015.

Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. Personalized ad recommendation systems for life-time value optimization with guarantees. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pp. 1806–1812. AAAI Press, 2015. ISBN 978-1-57735-738-4.

Towse, A., Jonsson, B., McGrath, C., Mason, A., Puig-Peiro, R., Mestre-Ferrandiz, J., Pistollato, M., and Devlin, N. Understanding variations in relative effectiveness: A health production approach. *International journal of technology assessment in health care*, 31(6):363–370, 2015.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Wooldridge, J. M. *Introductory econometrics: A modern approach*. Nelson Education, 2015.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. Conservative bandits. In *International Conference on Machine Learning*, pp. 1254–1262, 2016.