Robust Bandit Learning with Imperfect Context

Jianyi Yang, Shaolei Ren
University of California, Riverside
{jyang239, shaolei}@ucr.edu

Abstract

A standard assumption in contextual multi-arm bandit is that the true context is perfectly known before arm selection. Nonetheless, in many practical applications (e.g., cloud resource management), prior to arm selection, the context information can only be acquired by prediction subject to errors or adversarial modification. In this paper, we study a contextual bandit setting in which only imperfect context is available for arm selection while the true context is revealed at the end of each round. We propose two robust arm selection algorithms: MaxMinUCB (Maximize Minimum UCB) which maximizes the worst-case reward, and MinWD (Minimize Worst-case Degradation) which minimizes the worst-case regret. Importantly, we analyze the robustness of MaxMinUCB and MinWD by deriving both regret and reward bounds compared to an oracle that knows the true context. Our results show that as time goes on, MaxMinUCB and MinWD both perform as asymptotically well as their optimal counterparts that know the reward function. Finally, we apply MaxMinUCB and MinWD to online edge datacenter selection, and run synthetic simulations to validate our theoretical analysis.

Introduction

Contextual bandits (Lu, Pál, and Pál 2010; Chu et al. 2011) concern online learning scenarios such as recommendation systems (Li et al. 2010), mobile health (Lei, Tewari, and Murphy 2014), cloud resource provisioning (Chen and Xu 2019), wireless communications (Saxena et al. 2019), in which arms (a.k.a., actions) are selected based on the underlying context to balance the tradeoff between exploitation of the already learnt knowledge and exploration of uncertain arms (Auer et al. 2002; Auer, Cesa-Bianchi, and Fischer 2002; Bubeck and Cesa-Bianchi 2012; Dani et al. 2008).

The majority of the existing studies on contextual bandits (Chu et al. 2011; Valko et al. 2013; Saxena et al. 2019) assume that a perfectly accurate context is known before each arm selection. Consequently, as long as the agent learns increasingly more knowledge about reward, it can select arms with lower and lower average regrets. In many cases, however, the perfect (or true) context is not available to the agent prior to arm selection. Instead, the true context is revealed after taking an action at the end of each round (Kirschner and Krause 2019), but can be predicted using predictors, such as time series prediction(Brockwell et al. 2016; Gers, Schmidhuber, and Cummins 2000), to facilitate the agent's arm selection. For example, in wireless communications, the channel condition is subject to various attenuation effects (e.g., path loss and small-scale multi-path fading), and is critical context information for the transmitter configuration such as modulation and rate adaption (i.e., arm selection) (Goldsmith 2005; Saxena et al. 2019). But, the channel condition context is predicted and hence can only be coarsely known until the completion of transmission. For another example, the exact workload arrival rate is crucial context information for cloud resource management, but cannot be known until the workload actually arrives. Naturally, context prediction is subject to prediction errors. Moreover, it can also open a new attack surface — an outside attacker may adversarially modify the predicted context. For example, a recent study (Chen, Tan, and Zhang 2019) shows that the energy load predictor in smart grid can be adversarially attacked to produce load estimates with higher-than-usual errors. More motivating examples are provided in (Yang and Ren 2021). In general, imperfectly predicted and even adversarially presented context is very common in practice.

As motivated by practical problems, we consider a bandit setting where the agent receives imperfectly predicted context and selects an arm at the beginning of each round and the context is revealed after arm selection. We focus on robust arm optimization given imperfect context, which is as crucial as robust reward function estimation or exploration in contextual bandits (Dudík, Langford, and Li 2011; Neu and Olkhovskaya 2020; Zhu et al. 2018). Concretely, with imperfect context, our goal is to select arms online in a robust manner to optimize the worst-case performance in a neighborhood domain with the received imperfect context as center and a defense budget as radius. In this way, the robust arm selection can defend against the imperfect context error (from either context prediction error or adversarial modification) constrained by the budget.

Importantly and interestingly, given imperfect context, maximizing the worst-case reward (referred to as type-I robustness objective) and minimizing the worst-case regret (referred to as type-II robustness objective) can lead to different arms, while they are the same under the setting of perfect context (Saxena et al. 2019; Li et al. 2010; Slivkins

2019). Given imperfect context, the strategy for type-I robustness is more conservative than that for type-II robustness in terms of reward. The choice of the robustness objective depends on applications. For example, some safety-aware applications (Sun, Dey, and Kapoor 2017; Garcıa and Fernández 2015) intend to avoid extremely low reward, and thus type-I objective is suitable for them. Other applications (Li et al. 2010; Chen et al. 2018; Guan et al. 2020) focus on preventing large sub-optimality of selected arms, and type-II objective is more appropriate. As a distinction from other works on robust optimization of bandits (Bogunovic et al. 2018; Kirschner et al. 2020; Nguyen et al. 2020), we highlight the difference of the two types of robustness objectives.

We derive two algorithms — MaxMinUCB (Maximize Minimum UCB), which maximizes the worst-case reward for type-I objective, and MinWD (Minimize Worst-case Degradation), which minimizes the worst-case regret for type-II objective. The challenge of algorithm designs is that the agent has no access to exact knowledge of reward function but the estimated counterpart based on history collected data. Thus, in our design, MaxMinUCB maximizes the lower bound of reward, while MinWD minimizes the upper bound of regret.

We analyze the robustness of MaxMinUCB and MinWD by deriving both regret and reward bounds, compared to a strong oracle that knows the true context for arm selection as well as the exact reward function. Importantly, our results show that, while a linear regret term exists for both MaxMinUCB and MinWD due to imperfect context, the added linear regret term is actually the same as the amount of regret incurred by respectively optimizing type-I and type-II objectives with perfect knowledge of the reward function. This implies that as time goes on, MaxMinUCB and MinWD will asymptotically approach the corresponding optimized objectives from the reward and regret views, respectively.

Finally, we apply MaxMinUCB and MinWD to the problem of online edge datacenter selection and run synthetic simulations to validate our theoretical analysis.

Related Work

Contextual bandits. Linear contextual bandit learning is considered in LinUCB by (Li et al. 2010). . The study (Abbasi-Yadkori, Pál, and Szepesvári 2011) improves the regret analysis of linear contextual bandit learning, while the studies (Agrawal and Goyal 2012, 2013) solve this problem by Thompson sampling and give a regret bound. There are also studies to extend the algorithms to general reward functions like non-linear functions, for which kernel method is exploited in GP-UCB (Srinivas et al. 2010), Kernel-UCB (Valko et al. 2013), IGP-UCB and GP-TS (Chowdhury and Gopalan 2017; Deshmukh, Dogan, and Scott 2017). Nonetheless, a standard assumption in these studies is that perfect context is available for arm selection, whereas imperfect context is common in many practical applications (Kirschner et al. 2020).

Adversarial bandits and Robustness. The prior studies on adversarial bandits (Auer and Chiang 2016; Jun et al. 2018; Altschuler, Brunel, and Malek 2019; Liu and Shroff 2019) have primarily focused on that the adversary maliciously presents rewards to the agent or directly injects errors in rewards. Moreover, many studies (Audibert and Bubeck 2009; Gerchinovitz and Lattimore 2016) consider the best constant policy throughout the entire learning process as the oracle, while in our setting the best arm depends on the true context at each round. The adversarial setting has also been extended to contextual bandits (Neu and Olkhovskaya 2020; Syrgkanis, Krishnamurthy, and Schapire 2016; Han et al. 2020).

Recently, robust bandit algorithms have been proposed for various adversarial settings. Some focus on robust reward estimation and exploration (Altschuler, Brunel, and Malek 2019; Guan et al. 2020; Dudík, Langford, and Li 2011), and others train a robust or distributionally robust policy (Wu et al. 2016; Syrgkanis, Krishnamurthy, and Schapire 2016; Si et al. 2020b,a). Our study differs from the existing adversarial bandits by seeking two different robust algorithms given imperfect (and possibly adversarial) context.

Optimization and bandits with imperfect context. (Rakhlin and Sridharan 2013) considers online optimization with predictable sequences and (Jadbabaie et al. 2015) focuses on adaptive online optimization competing with dynamic benchmarks. Besides, (Chen et al. 2014; Jiang et al. 2013) study the robust optimization of mini-max regret. These studies assume perfectly known cost functions without learning. A recent study (Bogunovic et al. 2018) considers Bayesian optimization and aims at identifying a worst-case good input region with input perturbation (which can also model a perturbed but fixed environment/context parameter). The study (Wang, Wu, and Wang 2016) considers the linear bandit where certain context features are hidden, and uses iterative methods to estimate hidden contexts and model parameters. Another recent study (Kirschner and Krause 2019) assumes the knowledge of context distribution for arm selection, and considers a weak oracle that also only knows context distribution. The relevant papers (Kirschner et al. 2020) and (Nguyen et al. 2020) consider robust Bayesian optimizations where context distribution information is imperfectly provided, and propose to maximize the worst-case expected reward for distributional robustness. Although the objective of MaxMinUCB in our paper is similar to the robust optimization objectives in the two papers, we additionally derive a lower bound for the true reward in our analysis, which provides another perspective on the robustness of arm selection. More importantly, considering that the objectives in the two relevant papers are equivalent to minimizing a pseudo robust regret, we propose MinWD and derive an upper bound for the incurred true regret.

Problem Formulation

Assume that at the beginning of round t, the agent receives imperfect context $\hat{x}_t \in \mathcal{X}$ which is exogenously provided and not necessarily the true context x_t. Given the imperfect context $\hat{\mathbf{x}}_t \in \mathcal{X}$ and an arm set $\mathcal{A}$, the agent selects an arm $a_t \in \mathcal{A}$ for round t. Then, the reward y_t along with the true context x_t is revealed to the agent at the end of round t. Assume that $\mathcal{X} \times \mathcal{A} \subseteq \mathbb{R}^d$, and we use $x_{a_t,t}$ to denote the d-dimensional concatenated vector $[x_t, a_t]$.

The reward y_t received by the agent in round t is jointly decided by the true context x_t and selected arm a_t, and can be expressed as follows

$$y_t = f(x_t, a_t) + n_t, \tag{1}$$

where $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathcal{X}$ is the context domain, and n_t is the noise term. We assume that the reward function f belongs to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ generated by a kernel function $k : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$. In this RKHS, there exists a mapping function $\phi : (\mathcal{X} \times \mathcal{A}) \to \mathcal{H}$ which maps context and arm to their corresponding feature in $\mathcal{H}$. By reproducing property, we have $k([x,a],[x',a']) = \langle \phi(x,a), \phi(x',a') \rangle$ and $f(x,a) = \langle \phi(x,a), \theta \rangle$ where θ is the representation of function $f(\cdot,\cdot)$ in $\mathcal{H}$. Further, as commonly considered in the bandit literature (Slivkins 2019; Li et al. 2010), the noise n_t follows b-sub-Gaussian distribution for a constant $b \geq 0$, i.e. conditioned on the filtration $\mathcal{F}_{t-1} = \{x_\tau, y_{a,\tau}, a_\tau, \tau = 1, \cdots, t-1\}$, $\forall \sigma \in \mathbb{R}$, $\mathbb{E}[e^{\sigma n_t} | \mathcal{F}_{t-1}] \leq \exp\left(\frac{\sigma^2 b^2}{2}\right)$.

Without knowledge of reward function f, bandit algorithms are designed to decide an arm sequence $\{a_t, t = 1, \cdots, T\}$ to minimize the cumulative regret

$$R_T = \sum_{t=1}^{T} f(x_t, A^*(x_t)) - f(x_t, a_t), \tag{2}$$

where $A^*(x_t) = \arg\max_{a \in \mathcal{A}} f(x_t, a)$ is the oracle-optimal arm at round t given the true context x_t. When the received contexts are perfect, i.e. $\hat{x}_t = x_t$, minimizing the cumulative regret is equivalent to maximizing the cumulative reward $F_T = \sum_{t=1}^{T} f(x_t, a_t)$.

Context Imperfectness

The context error can come from a variety of sources, including imperfect context prediction algorithms and adversarial corruption (Kirschner et al. 2020; Chen, Tan, and Zhang 2019) on context. We simply use context error to encapsulate all the error sources without further differentiation. We assume that context error $\|x_t - \hat{x}_t\|$, where $\|\cdot\|$ is a certain norm (Bogunovic et al. 2018), is less than Δ. Also, Δ is referred to as the defense *budget* and can be considered as the level of robustness/safeguard that the agent intends to provide against context errors: with a larger Δ, the agent wants to make its arm selection robust against larger context errors (at the possible expense of its reward). A time-varying error budget can be captured by using Δ_t. Denote the neighborhood domain of context x as $\mathcal{B}_\Delta(x) = \{y \in \mathcal{X} \mid \|y - x\| \leq \Delta\}$. Then, we have the true context $x_t \in \mathcal{B}_\Delta(\hat{x}_t)$, where $\hat{x}_t$ is available to the agent.

Reward Estimation

Reward estimation is critical for arm selection. Kernel ridge regression, which is widely used in contextual bandits (Slivkins 2019) serves as the reward estimation method in our algorithm designs. By kernel ridge regression, the estimated reward given arm a and context x is expressed as

$$\hat{f}_t(x,a) = \mathbf{k}_t^T(x,a)(\mathbf{K}_t + \lambda \mathbf{I})^{-1} \mathbf{y}_t \tag{3}$$

where $\mathbf{I}$ is an identity matrix, $\mathbf{y}_t \in \mathbb{R}^{t-1}$ contains the history of y_τ, $\mathbf{k}_t(x,a) \in \mathbb{R}^{t-1}$ contains $k([x,a],[x_\tau,a_\tau])$, and $\mathbf{K}_t \in \mathbb{R}^{(t-1)\times(t-1)}$ contains $k([x_\tau,a_\tau],[x_{\tau'},a_{\tau'}])$, for $\tau,\tau' \in \{1,\cdots,t-1\}$.

The confidence width of kernel ridge regression is given in the following concentration lemma followed by a definition of reward UCB.

Lemma 1 (Concentration of Kernel Ridge Regression)**.** *Assume that the reward function $f(x,a)$ satisfies $|f(x,a)| \leq B$, the noise n_t satisfies a sub-Gaussian distribution with parameter b, and kernel ridge regression is used to estimate the reward function. With a probability of at least $1-\delta, \delta \in (0,1)$, for all $a \in \mathcal{A}$ and $t \in \mathbb{N}$, the estimation error satisfies $|\hat{f}_t(x,a) - f(x,a)| \leq h_t s_t(x,a)$, where $h_t = \sqrt{\lambda}B + b\sqrt{\gamma_t - 2\log(\delta)}$, $\gamma_t = \log\det(\mathbf{I} + \mathbf{K}_t/\lambda) \leq \bar{d}\log(1 + \frac{t}{\bar{d}\lambda})$ and $\bar{d}$ is the rank of $\mathbf{K}_t$. Let $\mathbf{V}_t = \lambda\mathbf{I} + \sum_{s=1}^{t} \phi(x,a)\phi(x,a)^\top$, the squared confidence width is given by $s_t^2(x,a) = \phi(x,a)^\top \mathbf{V}_{t-1}^{-1} \phi(x,a) = \frac{1}{\lambda}k([x,a],[x,a]) - \frac{1}{\lambda}\mathbf{k}_t(x,a)^T(\mathbf{K}_t + \lambda\mathbf{I})^{-1}\mathbf{k}_t(x,a)$.*

Definition 1. Given arm $a \in \mathcal{A}$ and context $x \in \mathcal{X}$, the reward UCB (Upper Confidence Bound) is defined as $U_t(x,a) = \hat{f}_t(x,a) + h_t s_t(x,a)$.

The next lemma shows that the term $s_t(x_t, a_t)$ has a vanishing impact on regret over time.

Lemma 2. *The sum of confidence widths given x_t for $t \in \{1,\cdots,T\}$ satisfies $\sum_{t=1}^{T} s_t^2(x_t,a_t) \leq 2\gamma_T$, where $\gamma_T = \log\det(\mathbf{I} + \mathbf{K}_T/\lambda) \leq \bar{d}\log(1 + \frac{T}{\bar{d}\lambda})$ and $\bar{d}$ is the rank of $\mathbf{K}_T$.*

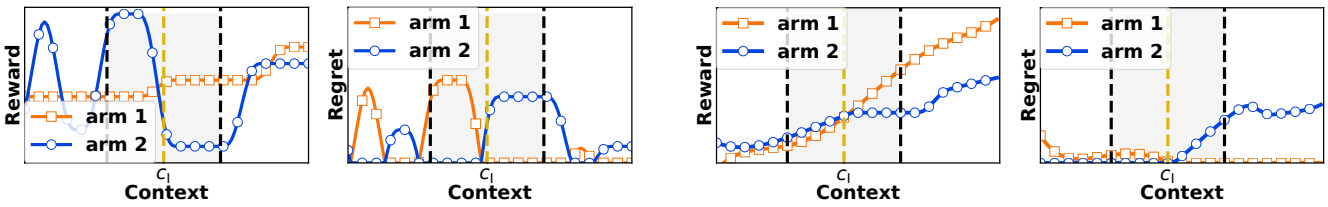
Then, we give the definition of *UCB-optimal* arm which is important in our algorithm designs.

Definition 2. Given context $x \in \mathcal{X}$, the *UCB-optimal* arm is defined as $A_t^\dagger(x) = \arg\max_{a \in \mathcal{A}} U_t(x,a)$.

Note that if the received contexts are perfect, i.e. $\hat{x}_t = x_t$, the standard contextual UCB strategy selects arm at round t as $A_t^\dagger(x_t)$. Under the cases with imperfect context, a naive policy (which we call SimpleUCB) is simply oblivious of context errors, i.e. the agent selects the UCB-optimal arm regarding imperfect context $\hat{x}_t$, denoted as $a_t^\dagger = A_t^\dagger(\hat{x}_t)$, by simply viewing the imperfect context $\hat{x}_t$ as true context. Nonetheless, if we want to guarantee the arm selection performance even in the worst case, robust arm selection that accounts for context errors is needed.

Robustness Objectives

In the existing bandit literature such as (Auer and Chiang 2016; Han et al. 2020; Li et al. 2010), maximizing the cumulative reward is equivalent to minimizing the cumulative regret, under the assumption of perfect context for arm selection. In this section, we will show that maximizing the worst-case reward is equivalent to minimizing a *pseudo* regret and is different from minimizing the worst-case true regret.



(a) Type-I Robustness (b) Type-II Robustness

Figure 1: Illustration of reward and regret functions that Type-I and Type-II robustness objectives are suitable for, respectively. The golden dotted vertical line represents the imperfect context c_{I}, and the gray region represents the defense region $\mathcal{B}_\Delta(c_{\mathrm{I}})$.

Type-I Robustness

With imperfect context, one approach to robust arm selection is to maximize the worst-case reward. With perfect knowledge of reward function, the oracle arm that maximizes the worst-case reward at round t is

$$\bar{a}_t = \arg\max_{a\in\mathcal{A}} \min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f(x,a). \tag{4}$$

For analysis in the following sections, given $\bar{a}_t$, the corresponding context for the worst-case reward is denoted as

$$\bar{x}_t = \arg\min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f\left(x,\bar{a}_t\right), \tag{5}$$

and the resulting optimal worst-case reward is denoted as

$$MF_t = f\left(\bar{x}_t,\bar{a}_t\right). \tag{6}$$

Next, Type-I robustness objective is defined based on the difference $\sum_{t=1}^{T} MF_t - F_T$, where $F_T = \sum_{t=1}^{T} f(x_t,a_t)$ is the actual cumulative reward.

Definition 3. If, with an arm selection strategy $\{a_1,\cdots,a_T\}$, the difference between the optimal cumulative worst-case reward and the cumulative true reward $\sum_{t=1}^{T} MF_t - F_T$ is sub-linear with respect to T, then the strategy achieves Type-I robustness.

If an arm selection strategy achieves Type-I robustness, the lower bound for the true reward $f\left(x_t,a_t\right)$ approaches the optimal worst-case reward MF_t in the defense region as t increases. Therefore, a strategy achieving type-I robustness objective can prevent very low reward. For example, in Fig. 1(a), arm 1 is the one that maximizes the worst-case reward, which is not necessarily optimal but always avoids extremely low reward under any context in the defense region.

Note that maximizing the worst-case reward is equivalent to minimizing the robust regret defined in (Kirschner et al. 2020), which is written using our formulation as

$$\bar{R}_T = \sum_{t=1}^{T} \min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f\left(x,\bar{a}_t\right) - \min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f\left(x,a_t\right). \tag{7}$$

However, this robust regret is a *pseudo* regret because the rewards of oracle arm $\bar{a}_t$ and selected arm a_t are compared under different contexts (i.e., their respective worst-case contexts), and it is not an upper or lower bound of the true regret R_T. To obtain a robust regret performance, we need to define another robustness objective based on the true regret.

Type-II Robustness

To provide robustness for the regret with imperfect context, we can minimize the cumulative worst-case regret, which is expressed as

$$\tilde{R}_T = \sum_{t=1}^{T} \max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} \left[f\left(x,A^*(x)\right) - f\left(x,a_t\right)\right]. \tag{8}$$

Clearly, the true regret $R_T \le \tilde{R}_T$, and minimizing the worst-case regret is equivalent to minimizing an upper bound for the true regret. Define the instantaneous regret function with respect to context x and arm a as $r\left(x,a\right) = f(x,A^*\left(x\right)) - f(x,a)$. Since given the reward function the optimization is decoupled among different rounds, the robust oracle arm to minimize the worst-case regret at round t is

$$\tilde{a}_t = \arg\min_{a\in\mathcal{A}} \max_{x\in\mathcal{B}_t(\hat{x}_t)} r(x,a). \tag{9}$$

For analysis in the following sections, given $\tilde{a}_t$, the corresponding context for the worst-case regret is denoted as

$$\tilde{x}_t = \arg\max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} r(x,\tilde{a}_t), \tag{10}$$

and the resulting optimal worst-case regret is

$$MR_t = r(\tilde{x}_t,\tilde{a}_t). \tag{11}$$

Now, we can give the definition of Type-II robustness as follows.

Definition 4. If, with an arm selection strategy $\{a_1,\cdots,a_T\}$, the difference between the cumulative true regret and the optimal cumulative worst-case regret $R_T - \sum_{t=1}^{T} MR_t$ is sub-linear with respect to T, then the strategy achieves Type-II robustness.

If an arm selection strategy achieves Type-II robustness, as time increases, the upper bound for the true regret $r(x_t,a_t)$ also approaches the optimal worst-case regret MR_t. Hence, a strategy achieving type-II robustness objective can prevent a high regret. As shown in Fig. 1(b), arm 1 is selected by minimizing the worst-case regret, which is a robust arm selection because the regret of arm 1 under any context in the defense region is not too high.

Comparison of Two Robustness Objectives

The two types of robustness correspond to the algorithms maximizing the worst-case reward and minimizing the

Algorithm 1 Robust Arm Selection with Imperfect Context

Input: Context error budget Δ
for $t = 1, \cdots, T$ **do**
Receive imperfect context $\hat{x}_t$.
Select arm a_t^{I} to solve Eqn. (12) in MaxMinUCB; or select arm a_t^{II} to solve Eqn. (16) in MinWD
Observe the true context x_t and the reward y_t.
end for

worst-case regret, respectively. In many cases, they result in different arm selections. Take the two scenarios in Fig. 1 as examples. In the scenario of Fig. 1(a), given the defense region, arm 1 is selected by maximizing the worst-case reward and arm 2 is selected by minimizing the worst-case regret. It can be observed that the worst-case regrets of the two arms are very close, but the worst-case reward of arm 2 is much lower than that of arm 1. Thus, the strategy of maximizing the worst-case reward is more suitable for this scenario. Differently, in the scenario of Fig. 1(b), arm 2 is selected by maximizing the worst-case reward and arm 1 is selected by minimizing the worst-case regret. Since the worst-case rewards of the two arms are very close and the worst-case regret of arm 2 is much larger than arm 1, it is more suitable to minimize the worst-case regret.

Robust Bandit Arm Selection

In this section, we propose two robust arm selection algorithms: (1) MaxMinUCB (Maximize Minimum Upper Confidence Bound), which aims to maximize the minimum reward (Type-I robustness objective); and (2) MinWD (Minimize Worst-case Degradation), which aims to minimize the maximum regret (Type-II robustness objective). We derive the regret and reward bounds for both algorithms and the proofs are available in (Yang and Ren 2021).

MaxMinUCB: Maximize Minimum UCB

Algorithm To achieve type-I robustness, MaxMinUCB in Algorithm 1 selects an arm a_t^{I} by maximizing the minimum UCB within the defense region $\mathcal{B}_\Delta(\hat{x}_t)$:

$$a_t^{\mathrm{I}} = \arg\max_{a\in\mathcal{A}} \min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} U_t(x,a). \tag{12}$$

The corresponding context that attains the minimum UCB in Eqn.(12) is $x_t^{\mathrm{I}} = \min_{x\in\mathcal{B}_\Delta(\hat{x}_t)} U_t\left(x, a_t^{\mathrm{I}}\right)$.

Analysis The next theorem gives a lower bound of the cumulative true reward of MaxMinUCB in terms of the optimal worst-case reward and a sub-linear term.

Theorem 3. *If* MaxMinUCB *is used to select arms with imperfect context, then for any true contexts $x_t \in \mathcal{B}_\Delta(\hat{x}_t)$ at round $t, t = 1, \cdots, T$, with a probability of $1-\delta, \delta \in (0,1)$, we have the following lower bound on the worst-case cumulative reward*

$$F_T \geq \sum_{t=1}^{T} MF_t - 2h_T\sqrt{2T\bar{d}\log(1+\frac{T}{\bar{d}\lambda})} \tag{13}$$

where MF_t is the optimal worst-case reward in Eqn. (6)*, $\bar{d}$ is the rank of $\mathbf{K}_t$ and h_T is given in Lemma 1.*

Remark 1. Theorem 3 shows that by MaxMinUCB, the difference between the optimal cumulative worst-case reward and the cumulative true reward is sub-linear and thus effectively achieves Type-I robustness according to Definition 3. This means that the reward by MaxMinUCB has a bounded sub-linear gap compared to the optimal worst-case reward $\sum_{t=1}^{T} MF_t$ obtained with perfect knowledge of the reward function. □

We are also interested in the cumulative true regret of MaxMinUCB which is given in the following corollary.

Corollary 3.1. *If* MaxMinUCB *is used to select arms with imperfect context, then for any true contexts $x_t \in \mathcal{B}_\Delta(\hat{x}_t)$ at round $t, t = 1, \cdots, T$, with a probability of $1-\delta, \delta \in (0,1)$, we have the following bound on the cumulative true regret defined in Eqn.* (2)*:*

$$R_T \leq \sum_{t=1}^{T} \overline{MR}_t + 2h_T\sqrt{2T\bar{d}\log(1+\frac{T}{\bar{d}\lambda})} \tag{14}$$

where $\overline{MR}_t = \max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f(x, A^*(x)) - MF_t$, MF_t is the optimal worst-case reward in Eqn. (6) *.*

Remark 2. Corollary 3.1 shows that the worst-case regret by MaxMinUCB can be quite larger than the optimal worst-case regret MR_t given in Eqn. (11) (Type-II robustness objective). Actually, despite being robust in terms of rewards, arms selected by MaxMinUCB can still have very large regret as shown in Fig. 1(b). Thus, to achieve type-II robustness, it is necessary to develop an arm selection algorithm that minimizes the worst-case regret.

MinWD: Minimize Worst-case Degradation

Algorithm MinWD is designed to asymptotically minimize the worst-case regret. Without the oracle knowledge of reward function, MinWD performs arm selection based on the upper bound of regret. Denote $D_a(x) = U_t\left(x, A_t^\dagger(x)\right) - U_t(x,a)$ referred to as UCB *degradation* at context x. By Lemma 1, the instantaneous true regret can be bounded as

$$\begin{aligned} r(x_t,a_t) &\leq [D_{a_t}(x_t) + 2h_t s_t(x_t,a_t)] \\ &\leq \overline{D}_{a_t} + 2h_t s_t(x_t,a_t), \end{aligned} \tag{15}$$

where $\overline{D}_{a_t} = \max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} D_a(x)$ is called the *worst case degradation*, and $2h_t s_t(x_t,a_t)$ has a vanishing impact by Lemma 2. Thus, to minimize worst-case regret, MinWD minimizes its upper bound $\overline{D}_{a_t}$ excluding the vanishing term $2h_t s_t(x_t,a_t)$, i.e.

$$a_t^{\mathrm{II}} = \min_{a\in\mathcal{A}} \max_{\mathbf{x}\in\mathcal{B}_\Delta(\hat{x}_t)} \left\{U_t\left(x, A_t^\dagger(x)\right) - U_t(x,a)\right\}. \tag{16}$$

The context that attains the worst case in Eqn. (16) is written as $x_t^{\mathrm{II}} = \arg\max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} D_{a_t^{\mathrm{II}}}(x)$.

Analysis Given arm a_t^{II} selected by MinWD, the next lemma gives an upper bound of worst-case degradation.

Lemma 4. *If* MinWD *is used to select arms with imperfect context, then for each $t = 1, 2, \cdots, T$, with a probability at least $1-\delta, \delta \in (0,1)$, we have*

$$\overline{D}_{a_t^{\mathrm{II}},t} \leq MR_t + 2h_t s_t\left(\dot{x}_t, A_t^{\dagger}\left(\dot{x}_t\right)\right), \tag{17}$$

where MR_t is the optimal worst-case regret defined in Eqn. (11)*, $\dot{x}_t = \arg\max_{x \in \mathcal{B}_{\Delta}(\hat{x}_t)} D_{\tilde{a}_t}(x)$ is the context that maximizes the degradation given the arm $\tilde{a}_t$ defined for the optimal worst-case regret in Eqn.* (10)*.*

Then, in order to show that $\overline{D}_{a_t^{\mathrm{II}},t}$ approaches MR_t, we need to prove that $2h_t s_t\left(\dot{x}_t, A_t^{\dagger}\left(\dot{x}_t\right)\right)$ vanishes as t increases. But, this is difficult because the considered sequence $\left\{\dot{x}_t, A_t^{\dagger}\left(\dot{x}_t\right)\right\}$ is different from the actual sequence of context and selected arms $\left\{x_t, a_t^{\mathrm{II}}\right\}$ under MinWD. To circumvent this issue, we first introduce the concept of $\epsilon-$ covering (Wu 2016). Denote $\Phi = \mathcal{X} \times \mathcal{A}$ as the context-arm space. If a finite set Φ_ϵ is an $\epsilon-$ covering of the space Φ, then for each $\varphi \in \Phi$, there exists at least one $\bar{\varphi} \in \Phi_\epsilon$ satisfying $\|\varphi - \bar{\varphi}\|_2 \leq \epsilon$. Denote $\mathcal{C}_\epsilon(\bar{\varphi}) = \{\varphi \mid \|\varphi - \bar{\varphi}\|_2 \leq \epsilon\}$ as the cell with respect to $\bar{\varphi} \in \Phi_\epsilon$. Since the dimension of the entries in Φ is d, the size of the Φ_ϵ is $|\Phi_\epsilon| \sim O\left(\frac{1}{\epsilon^d}\right)$. Besides, we assume the mapping function ϕ is Lipschitz continuous, i.e. $\forall x, y \in \Phi$, $\|\phi(x) - \phi(y)\| \leq L_\phi \|x - y\|$. Next, we prove the following proposition to bound the sum of confidence widths under some conditions.

Proposition 5. *Let $\mathcal{X}_T = \{x_{a_1,1}, \cdots, x_{a_T,T}\}$ be the sequence of true contexts and selected arms by bandit algorithms and $\dot{\mathcal{X}}_T = \{\dot{x}_{\dot{a}_1,1}, \cdots, \dot{x}_{\dot{a}_T,T}\}$ be the considered sequence of contexts and actions. Suppose that both $x_{a_t,t}$ and $\dot{x}_{\dot{a}_t,t}$belong to Φ. Besides, with an $\epsilon-$ covering $\Phi_\epsilon \subseteq \Phi$, $\epsilon > 0$, there exists $\kappa \geq 0$ such that two conditions are satisfied: First, $\forall \bar{\varphi} \in \Phi_\epsilon$, $\exists t \leq \lceil \kappa/\epsilon^d \rceil$ such that $x_{a_t,t} \in \mathcal{C}_\epsilon(\bar{\varphi})$. Second, if at round t, $x_{a_t,t} \in \mathcal{C}_\epsilon(\bar{\varphi})$ for some $\bar{\varphi} \in \Phi_\epsilon$, then $\exists t \leq t' < t + \lceil \kappa/\epsilon^d \rceil$ such that $x_{a'_{t},t'} \in \mathcal{C}_\epsilon(\bar{\varphi})$. If the mapping function ϕ is Lipschitz continuous with constant L_ϕ, the sum of squared confidence widths is bounded as*

$$\sum_{t=1}^{T} s_t^2\left(\dot{x}_{\dot{a}_t,t}\right) \leq \sqrt{T}\left(4\tilde{d}\log\left(1+\frac{T}{\tilde{d}\lambda}\right)+\frac{1}{\lambda}\right)+\frac{8L_\phi^2\kappa^{2/d}}{\lambda}T^{1-1/d},$$

where d is the dimension of $x_{a_t,t}$, $\tilde{d}$ is the effective dimension defined in the proof, $s_t^2\left(\dot{x}_{\dot{a}_t,t}\right) = \phi(\dot{x}_{\dot{a}_t,t})^\top \mathbf{V}_{t-1}^{-1}\phi(\dot{x}_{\dot{a}_t,t})$ and $\mathbf{V}_t = \lambda\mathbf{I} + \sum_{s=1}^{t}\phi(x_{a_s,s})\phi(x_{a_s,s})^\top$.

Remark 3. The conditions in Proposition 5 guarantee that the time interval between the events that true context-arm feature lies in the same cell is not larger than $\lceil \kappa/\epsilon^d \rceil$, which is proportional to the size of the ϵ-covering $|\Phi_\epsilon|$. That means, similar contexts and selected arms occur in the true sequence repeatedly if T is large enough. If contexts are sampled from a bounded space $\mathcal{X}$ with some distribution, then similar contexts will occur repeatedly. Also, note that the arm in our considered sequence $A_t^{\dagger}\left(\dot{x}_t\right)$ is the UCB-optimal arm, which becomes close to the optimal arm for $\dot{x}_t$ if the confidence width is sufficiently small. Hence, there exists some context error budget sequence $\{\Delta_t\}$ such that, starting from a certain round T_0, the two conditions are satisfied. The two conditions in Proposition 5 are mainly for theoretical analysis of MinWD.

By Lemma 4 and Proposition 5, we bound the cumulative regret of MinWD.

Theorem 6. *If* MinWD *is used to select arms with imperfect context and as time goes on, and the conditions in Proposition 5 are satisfied, then for any true context $x_t \in \mathcal{B}_\Delta(\hat{x}_t)$ at round $t, t = 1, \cdots, T$, with a probability of $1-\delta, \delta \in (0,1)$, we have the following bound on the cumulative true regret:*

$$R_T \leq \sum_{t}^{T} MR_t + 2h_T T^{\frac{3}{4}}\sqrt{\left(4\tilde{d}\log\left(1+\frac{T}{\tilde{d}\lambda}\right)+\frac{1}{\lambda}\right)}+ 4\sqrt{\frac{2}{\lambda}}L_\phi\kappa^{\frac{1}{d}}h_T T^{1-\frac{1}{2d}} + 2h_T\sqrt{2T\bar{d}\log(1+\frac{T}{\bar{d}\lambda})},$$

where MR_t is the optimal worst-case regret for round t in Eqn. (11)*, d is the dimension of $x_{a_t,t}$, $\tilde{d}$ is the effective dimension defined in the proof of Proposition 5, $\bar{d}$ is the rank of $\mathbf{K}_t$ and h_T is given in Lemma 1.*

Remark 4. Theorem 6 shows that by MinWD, $R_T - \sum_{t=1}^{T} MR_t$ is sub-linear w.r.t. T and thus Type-II robustness is effectively achieved according to Definition 4. This means the true regret bound approaches $\sum_t^T MR_t$, the optimal worst-case regret, asymptotically.

Next, in parallel with MaxMinUCB, we derive the bound of true reward for MinWD.

Corollary 6.1. *If* MinWD *is used to select arms with imperfect context and as time goes on, and the true sequence of context and arm obeys the conditions in Proposition 5, then for any true contexts $x_t \in \mathcal{B}_\Delta(\hat{x}_t)$ at round $t, t = 1, \cdots, T$, with a probability of $1-\delta, \delta \in (0,1)$, we have the following lower bound of the cumulative reward*

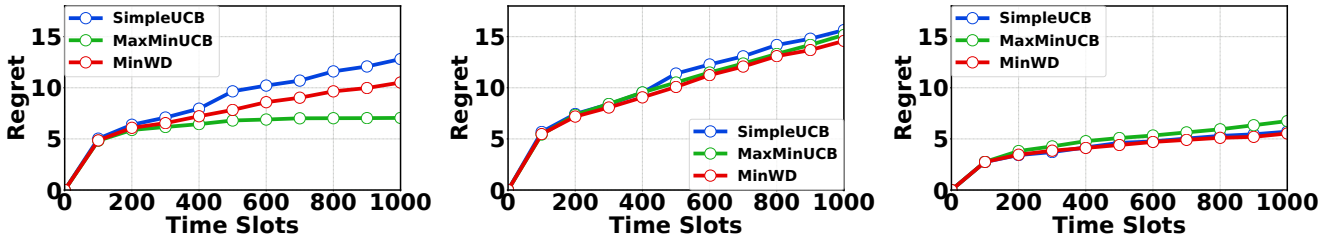
$$F_T \geq \sum_{t=1}^{T}[MF_t - MR_t] - 2h_T T^{\frac{3}{4}}\sqrt{\left(4\tilde{d}\log\left(1+\frac{T}{\tilde{d}\lambda}\right)+\frac{1}{\lambda}\right)} - 4\sqrt{\frac{2}{\lambda}}L_\phi\kappa^{\frac{1}{d}}h_T T^{1-\frac{1}{2d}} - 2h_T\sqrt{2T\bar{d}\log(1+\frac{T}{\bar{d}\lambda})},$$

where MR_t is the optimal worst-case regret for round t in Eqn. (11)*, d is the dimension of $x_{a_t,t}$, $\tilde{d}$ is the effective dimension defined in the proof of Proposition 5, $\bar{d}$ is the rank of $\mathbf{K}_t$, and h_T is given in Lemma 1.*

Remark 5. Corollary 6.1 shows that as t becomes sufficiently large, the difference between the optimal worst-case reward and the true reward of the selected arm is no larger than the optimal worst-case regret MR_t. With perfect context, we have $MR_t = 0$, and hence MaxMinUCB and MinWD both asymptotically maximize the reward, implying that these two types of robustness are the same under perfect context.

Summary of Main Results

We summarize our analysis of MaxMinUCB and MinWD in Table 1, while the algorithms details are available in Algorithm 1. In the table, d is the dimension of context-arm



(a) Robust regret. (b) Worst-case regret. (c) True regret.

Figure 2: Different cumulative regret objectives for different algorithms.

Algorithms	Regret	Reward
MaxMinUCB	$\sum_{t=1}^{T}\overline{MR}_t + O(\sqrt{T\log T})$	$\sum_{t=1}^{T}MF_t - O(\sqrt{T\log T})$
MinWD	$\sum_{t=1}^{T}MR_t + O(T^{\frac{3}{4}}\sqrt{\log T} + T^{1-\frac{1}{2d}}+\sqrt{T\log T})$	$\sum_{t}^{T}[MF_t - MR_t] + O(T^{\frac{3}{4}}\sqrt{\log T} + T^{1-\frac{1}{2d}}+\sqrt{T\log T})$

Table 1: Summary of Analysis

vector $[x, a]$, $\overline{MR}_t = \max_{x\in\mathcal{B}_\Delta(\hat{x}_t)} f(x, A^*(x)) - MF_t$, and MF_t and MR_t are defined in Eqn. (6) and (11), respectively. Type-I and type-II robustness objectives are achieved by MaxMinUCB and MinWD respectively.

Simulation

Edge computing is a promising technique to meet the demand of latency-sensitive applications (Shi et al. 2016). Given multiple heterogeneous edge datacenters located in different locations, which one should be selected? Specifically, each edge datacenter is viewed as an arm, and the users' workload is context that can only be predicted prior to arm selection. Our goal is to learn datacenter selection to optimize the latency in a robust manner given imperfect workload information. We assume that the service rate of the edge datacenter a, $a \in \mathcal{A}$, is μ_a, the computation latency satisfies an M/M/1 queueing model and the average communication delay between this datacenter and users is p_a. Hence, the average total latency cost can be expressed as $l(x, a) = p_a \cdot x + \frac{x}{\mu_a - x}$ which is commonly-considered in the literature (Lin et al. 2011; Xu, Chen, and Ren 2017; Lin et al. 2012). The detailed settings are given in (Yang and Ren 2021).

In Fig. 2, we compare different algorithms in terms of three cumulative regret objectives: robust regret in Eqn. (7), worst-case regret in Eqn. (8) and true regret in Eqn. (2). We consider the following algorithms: SimpleUCB with imperfect context, MaxMinUCB with imperfect context and MinWD with imperfect context. Given a sequence of true contexts, imperfect context sequence is generated by sampling i.i.d. uniform distribution over $\mathcal{B}_\Delta(x_t)$ at each round. In the simulations, Gaussian kernel with parameter 0.1 is used for reward (loss) estimation. λ in Eqn. (3) is set as 0.1. The exploration rate is set as $h_t = 0.04$.

As is shown in Fig. 2(a), MaxMinUCB has the best performance of robust regret among the three algorithms. This is because MaxMinUCB targets at type-I robustness objective which is equivalent to minimizing the robust regret. However, MaxMinUCB is not the best algorithm in terms of true regret as is shown in Fig. 2(c) since robust regret is not an upper or lower bound of true regret. Another robust algorithm MinWD is also better than SimpleUCB in terms of robust regret, and it has the best performance among the three algorithms in terms of the worst-case regret, as shown in Fig. 2(b). This is because the regret of MinWD approaches the optimal worst-case regret (Theorem 6). MinWD also has a good performance of true regret, which coincides with the fact that the worst-case regret is the upper bound of the true regret. By comparing the three algorithms in terms of the three regret objectives, we can clearly see that MaxMinUCB and MinWD achieve performance robustness in terms of the robust regret and worst-case regret, respectively.

Conclusion

In this paper, considering a bandit setting with imperfect context, we propose: MaxMinUCB which maximizes the worst-case reward; and MinWD which minimizes the worst-case regret. Our analysis of MaxMinUCB and MinWD based on regret and reward bounds shows that as time goes on, MaxMinUCB and MinWD both perform as asymptotically well as their counterparts that have perfect knowledge of the reward function. Finally, we consider online edge datacenter selection and run synthetic simulations for evaluation.

Acknowledgments

This work was supported in part by the NSF under grants CNS-1551661 and ECCS-1610471.

References

Abbasi-Yadkori, Y.; Pál, D.; and Szepesvári, C. 2011. Improved Algorithms for Linear Stochastic Bandits. *NeurIPS* .

Agrawal, S.; and Goyal, N. 2012. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. *COLT* .

Agrawal, S.; and Goyal, N. 2013. Thompson Sampling for Contextual Bandits with Linear Payoffs. *ICML* .

Altschuler, J.; Brunel, V.-E.; and Malek, A. 2019. Best Arm Identification for Contaminated Bandits. *Journal of Machine Learning Research* 20(91): 1–39.

Audibert, J.; and Bubeck, S. 2009. Minimax policies for adversarial and stochastic bandits. *COLT* .

Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47: 235–256.

Auer, P.; Cesa-Bianchi, N.; Freund, Y.; and Schapire, R. E. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing* 32: 48–77.

Auer, P.; and Chiang, C.-K. 2016. An Algorithm with Nearly Optimal Pseudo-regret for Both Stochastic and Adversarial Bandits. In *COLT*.

Bogunovic, I.; Scarlett, J.; Jegelka, S.; and Cevher, V. 2018. Adversarially Robust Optimization with Gaussian Processes. In *NIPS*.

Brockwell, P. J.; Brockwell, P. J.; Davis, R. A.; and Davis, R. A. 2016. *Introduction to time series and forecasting*. Springer.

Bubeck, S.; and Cesa-Bianchi, N. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5: 1–122.

Chen, B.; Wang, J.; Wang, L.; He, Y.; and Wang, Z. 2014. Robust optimization for transmission expansion planning: Minimax cost vs. minimax regret. *IEEE Transactions on Power Systems* 29(6): 3069–3077.

Chen, L.; and Xu, J. 2019. Budget-constrained edge service provisioning with demand estimation via bandit learning. *IEEE Journal on Selected Areas in Communications* 37(10): 2364–2376.

Chen, L.; Xu, J.; Ren, S.; and Zhou, P. 2018. Spatio–temporal edge service placement: A bandit learning approach. *IEEE Transactions on Wireless Communications* 17(12): 8388–8401.

Chen, Y.; Tan, Y.; and Zhang, B. 2019. Exploiting Vulnerabilities of Load Forecasting Through Adversarial Attacks. In *e-Energy*.

Chowdhury, S. R.; and Gopalan, A. 2017. On kernelized multi-armed bandits. *ICML* .

Chu, W.; Li, L.; Reyzin, L.; and Schapire, R. 2011. Contextual bandits with linear payoff functions. *NeurIPS* .

Dani, V.; Hayes, T.; Thomas, P.; and Kakade, S. 2008. Stochastic linear optimization under bandit feedback. *COLT* .

Deshmukh, A. A.; Dogan, U.; and Scott, C. 2017. Multi-Task Learning for Contextual Bandits. *NeurIPS* .

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601* .

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16(1): 1437–1480.

Gerchinovitz, S.; and Lattimore, T. 2016. Refined lower bounds for adversarial bandits. *NeurIPS* .

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to Forget: Continual Prediction with LSTM. *Neural Computation* 12(10): 2451–2471.

Goldsmith, A. 2005. *Wireless Communications*. Cambridge University Press.

Guan, Z.; Ji, K.; Bucci Jr, D. J.; Hu, T. Y.; Palombo, J.; Liston, M.; and Liang, Y. 2020. Robust Stochastic Bandit Algorithms under Probabilistic Unbounded Adversarial Attack. In *AAAI*.

Han, Y.; Zhou, Z.; Zhou, Z.; Blanchet, J.; Glynn, P. W.; and Ye, Y. 2020. Sequential Batch Learning in Finite-Action Linear Contextual Bandits. *arXiv preprint arXiv:2004.06321* .

Jadbabaie, A.; Rakhlin, A.; Shahrampour, S.; and Sridharan, K. 2015. Online optimization: Competing with dynamic comparators. In *AISTATS*.

Jiang, R.; Wang, J.; Zhang, M.; and Guan, Y. 2013. Two-stage minimax regret robust unit commitment. *IEEE Transactions on Power Systems* 28(3): 2271–2282.

Jun, K.-S.; Li, L.; Ma, Y.; and Zhu, X. 2018. Adversarial Attacks on Stochastic Bandits. In *NIPS*.

Kirschner, J.; Bogunovic, I.; Jegelka, S.; and Krause, A. 2020. Distributionally Robust Bayesian Optimization. In *AISTATS*.

Kirschner, J.; and Krause, A. 2019. Stochastic Bandits with Context Distributions. In *NeurIPS*.

Lei, H.; Tewari, A.; and Murphy, S. 2014. An actor-critic contextual bandit algorithm for personalized interventions using mobile devices. *NeurIPS* .

Li, L.; Chu, W.; Langford, J.; and Schapire, R. E. 2010. A Contextual-bandit Approach to Personalized News Article Recommendation. In *WWW*.

Lin, M.; Liu, Z.; Wierman, A.; and Andrew, L. L. H. 2012. Online algorithms for geographical load balancing. In *IGCC*.

Lin, M.; Wierman, A.; Andrew, L. L. H.; and Thereska, E. 2011. Dynamic right-sizing for power-proportional data centers. In *INFOCOM*.

Liu, F.; and Shroff, N. 2019. Data Poisoning Attacks on Stochastic Bandits. In *ICML*.

Lu, T.; Pál, D.; and Pál, M. 2010. Contextual multi-armed bandits. *AISTATS* .

Neu, G.; and Olkhovskaya, J. 2020. Efficient and robust algorithms for adversarial linear contextual bandits. In *COLT*.

Nguyen, T.; Gupta, S.; Ha, H.; Rana, S.; and Venkatesh, S. 2020. Distributionally robust bayesian quadrature optimization. In *AISTATS*.

Rakhlin, A.; and Sridharan, K. 2013. Online Learning with Predictable Sequences. In *COLT*.

Saxena, V.; Jaldén, J.; Gonzalez, J. E.; Bengtsson, M.; Tullberg, H.; and Stoica, I. 2019. Contextual Multi-Armed Bandits for Link Adaptation in Cellular Networks. In *Workshop on Network Meets AI & ML (NetAI)*.

Shi, W.; Cao, J.; Zhang, Q.; Li, Y.; and Xu, L. 2016. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal* 3(5): 637–646.

Si, N.; Zhang, F.; Zhou, Z.; and Blanchet, J. 2020a. Distributional Robust Batch Contextual Bandits. *arXiv preprint arXiv:2006.05630* .

Si, N.; Zhang, F.; Zhou, Z.; and BlanchetWu, J. 2020b. Distributionally Robust Policy Evaluation and Learning in Offline Contextual Bandits. In *ICML*.

Slivkins, A. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends in Machine Learning* 12(1-2): 1–286.

Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. Gaussian process optimization in the bandit setting: no regret and experimental design. *ICML* .

Sun, W.; Dey, D.; and Kapoor, A. 2017. Safety-aware algorithms for adversarial contextual bandit. In *ICML*.

Syrgkanis, V.; Krishnamurthy, A.; and Schapire, R. 2016. Efficient algorithms for adversarial contextual learning. *ICML* .

Valko, M.; Korda, N.; Munos, R.; Flaounas, I.; and Cristianini, N. 2013. Finite-time analysis of kernelised contextual bandits. *UAI* .

Wang, H.; Wu, Q.; and Wang, H. 2016. Learning Hidden Features for Contextual Bandits. *CIKM* .

Wu, Y. 2016. Packing, covering, and consequences on minimax risk. http://www.stat.yale.edu/~yw562/teaching/598/lec14.pdf.

Wu, Y.; Shariff, R.; Lattimore, T.; and Szepesvári, C. 2016. Conservative bandits. In *ICML*.

Xu, J.; Chen, L.; and Ren, S. 2017. Online Learning for Offloading and Autoscaling in Energy Harvesting Mobile Edge Computing. *IEEE Transactions on Cognitive Communications and Networking* 3(3): 361–373.

Yang, J.; and Ren, S. 2021. Robust Bandit Learning with Imperfect Context. *arXiv preprint arXiv:2102.05018* .

Zhu, F.; Guo, J.; Li, R.; and Huang, J. 2018. Robust actor-critic contextual bandit for mobile health (mhealth) interventions. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 492–501.