Robust Bandit Learning with Imperfect Context

Jianyi Yang, Shaolei Ren
University of California, Riverside
{jayang239, 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.

1 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; García 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.

## 2 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 Cesa-Bianchi 2002, 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 Valko 2018; 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; Dudik, 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. 2020). 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 out 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.

## 3 Problem Formulation

Assume that at the beginning of round \( t \), the agent receives imperfect context \( x_t \in X \) which is exogenously provided and not necessarily the true context \( x_t \). Given the imperfect context \( x_t \) and an arm set \( A \), the agent selects an arm \( a_t \in A \) for round \( t \). Then, the reward \( r_t \) along with the true context \( x_t \) is revealed to the agent at the end of round \( t \). Assume that \( X \times A \subseteq \mathbb{R}^d \), 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

$$y_t = f(x_t, a_t) + n_t,$$  

where $f : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$ 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 $k : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \rightarrow \mathbb{R}$. In this RKHS, there exists a mapping function $\phi : (\mathcal{X} \times \mathcal{A}) \rightarrow \mathcal{H}$ 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 $\theta$ is the representation of function $f(\cdot, \cdot)$ in $\mathcal{H}$. Further, as commonly considered in the bandit literature [Sivkins 2019, Li et al. 2010], the noise $n_t$ follows a sub-Gaussian distribution for a constant $b \geq 0$, i.e. conditioned on the filtration $\mathcal{F}_{t-1} = \{x_t, y_{0:t}, \theta, t = 1, \ldots, t-1\}, \forall \sigma \in \mathbb{R}$, $E[e^{\sigma n_t} | \mathcal{F}_{t-1}] \leq \exp \left( \frac{\sigma^2 b^2}{2} \right)$.

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

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

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. $\tilde{x}_t = x_t$, minimizing the cumulative regret is equivalent to maximizing the cumulative reward $\hat{R}_T = \sum_{t=1}^{T} f(x_t, a_t)$.

### 3.1 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 $\|x_t - \tilde{x}_t\|$, where $\| \cdot \|$ is a certain norm (Bogunovic et al. 2018), is less than $\Delta$. Also, $\Delta$ 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 $\Delta$, 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 $\Delta_t$. Denote the neighborhood domain of context $x$ as $B_\Delta(x) = \{ y \in \mathcal{X} \mid \|y - x\| \leq \Delta \}$. Then, we have the true context $x_t \in B_\Delta(\hat{x}_t)$, where $\hat{x}_t$ is available to the agent.

### 3.2 Reward Estimation

Reward estimation is critically important for arm selection. Kernel ridge regression, which is widely used in contextual bandits (Sivkins 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

$$\hat{f}(x, a) = k^T_t(x, a)(K_t + \lambda I)^{-1}y_t$$

where $I$ is an identity matrix, $y_t \in \mathbb{R}^{t-1}$ contains the history of $y_t$, $K_t(x, a) \in \mathbb{R}^{t-1}$ contains $k([x, a], [x, a])$, and $K_t \in \mathbb{R}^{(t-1) \times (t-1)}$ contains $k([x, a], [x, a])$, for $t, \tau \in \{1, \cdots, t-1\}$.

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 $h_t = \sqrt{b + \sqrt{\gamma_t - 2 \log(\delta)}}$, $\gamma_t = \log \det(I + K_t/\lambda) \leq d \log(1 + 1/d)$ and $d$ is the rank of $K_t$. Let $V_t = \lambda I + \sum_{s=1}^{t-1} \phi(x, a)\phi(x, a)^{\top}$, the squared confidence width is given by $s^2_t(x, a) = \phi(x, a)^{\top}V^{-1}_t \phi(x, a) = \frac{1}{\lambda} k([x, a], [x, a]) - \frac{1}{\lambda} k([x, a], [x, a])^{\top}(K_t + \lambda I)^{-1}k([x, a], [x, a])$.

**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 $\sum_{t=1}^{T} s^2_t(x_t, a_t) \leq 2\gamma T$, where $\gamma = \log \det(I + K_T/\lambda) \leq d \log(1 + \frac{1}{d})$ and $d$ is the rank of $K_T$.

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 $A^*_t(x) = \arg \max_{a \in \mathcal{A}} U_t(x, a)$.

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 $A^*_t(x_t)$. 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 $a^*_t = A^*_t(\hat{x}_t)$, 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.

### 4 Robustness Objectives

In the existing bandit literature such as [Auer and Cianetti 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.
### 4.1 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

$$\tilde{a}_t = \arg\max_{a \in A} \min_{x \in B_{\Delta}(\tilde{x}_t)} f(x, a).$$  \hfill (4)

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

$$\tilde{x}_t = \arg\min_{x \in B_{\Delta}(\tilde{x}_t)} f(x, \tilde{a}_t),$$  \hfill (5)

and the resulting optimal worst-case reward is denoted as

$$MF_t = f(\tilde{x}_t, \tilde{a}_t).$$  \hfill (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 $f(\tilde{x}_t, \tilde{a}_t)$ 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

$$\tilde{R}_T = \sum_{t=1}^{T} \min_{x \in B_{\Delta}(\tilde{x}_t)} f(x, \tilde{a}_t) - \min_{x \in B_{\Delta}(\tilde{x}_t)} f(x, a_t).$$  \hfill (7)

However, this robust regret is a pseudo regret because the rewards of oracle arm $\tilde{a}_t$ 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.

### 4.2 Type-II Robustness

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

$$\tilde{R}_T = \sum_{t=1}^{T} \max_{x \in B_{\Delta}(\tilde{x}_t)} [f(x, A^*(x)) - f(x, a_t)].$$  \hfill (8)

Clearly, the true regret $R_T \leq \tilde{R}_T$, 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 $r(x, a) = f(x, A^*(x)) - f(x, a)$. 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

$$\tilde{a}_t = \arg\min_{a \in A} \max_{x \in B_{\Delta}(\tilde{x}_t)} r(x, a).$$  \hfill (9)

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

$$\tilde{x}_t = \arg\max_{x \in B_{\Delta}(\tilde{x}_t)} r(x, \tilde{a}_t),$$  \hfill (10)

and the resulting optimal worst-case regret is

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

### 4.3 Comparison of Two Robustness Objectives

The two types of robustness correspond to the algorithms maximizing the worst-case reward and minimizing the
The next theorem gives a lower bound of the cumulative reward (Type-I robustness objective). We derive the regret and reward bounds for both algorithms and the proofs are available in (Yang and Ren 2021).

5 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).

5.1 MaxMinUCB: Maximize Minimum UCB

Algorithm To achieve type-I robustness, MaxMinUCB in Algorithm 1 selects an arm  at by maximizing the minimum UCB within the defense region  (, ):

\[
\hat{a}_t = \arg \max_{a \in A} \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a). \tag{12}
\]

The corresponding context that attains the minimum UCB in Eqn. (12) is \( x_t^I = \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a_t^I) \).

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 \( x_t \in B_\Delta(\hat{x}_t) \) at round \( t = 1, \ldots, T \), with a probability of \( 1 - \delta, \delta \in (0, 1) \), we have the following lower bound on the worst-case cumulative reward

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

where \( M F_t \) is the optimal worst-case reward in Eqn. (6), \( \bar{d} \) is the rank of \( K_t \) and \( h_T \) is given in Lemma 7.

**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 \( \sum_{t=1}^T M F_t \) 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 \( x_t \in B_\Delta(\hat{x}_t) \) at round \( t = 1, \ldots, T \), with a probability of \( 1 - \delta, \delta \in (0, 1) \), we have the following bound on the cumulative true regret defined in Eqn. (2):

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

where \( \overline{M R}_t = \max_{x \in B_\Delta(\hat{x}_t)} \{ f(x, A^*_t(x)) - M F_t \} \) is the optimal worst-case reward in Eqn. (2).

**Remark 2.** Corollary 3.1 shows that the worst-case regret by MaxMinUCB can be quite larger than the optimal worst-case regret \( M R_t \) 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.

5.2 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 \( D_a(x) = U_t(x, A_t^I(x)) - U_t(x, a) \) referred to as UCB degradation at context \( x \). By Lemma 4 the instantaneous true regret can be bounded as

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

where \( \overline{D}_{a_t} = \max_{x \in B_\Delta(\hat{x}_t)} D_a(x) \) is called the worst case degradation, and \( 2h_t s_t(x_t, a_t) \) has a vanishing impact by Lemma 3. 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.

\[
a_t^H = \min_{a \in A} \max_{x \in B_\Delta(\hat{x}_t)} \left\{ U_t(x, A_t^I(x)) - U_t(x, a) \right\}. \tag{16}
\]

The context that attains the worst case in Eqn. (16) is written as \( x_t^H = \arg \max_{x \in B_\Delta(\hat{x}_t)} D_{a_t^H}(x) \).

**Analysis** Given arm \( a_t^H \) 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

\[
\mathcal{D}_{a^i_{t,t}} \leq M R_t + 2h_t s_t \left( \hat{x}_t, A^i_t (\hat{x}_t) \right),
\]

where \( M \) is the optimal worst-case regret defined in Eqn. (11), \( \hat{x}_t = \arg \max_{x \in \mathcal{X}_t} \mathcal{D}_{a^i_t} (x) \) 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 \( \mathcal{D}_{a^i_{t,t}} \) approaches \( M R_t \), we need to prove that \( 2h_t s_t \left( \hat{x}_t, A^i_t (\hat{x}_t) \right) \) vanishes as \( t \) increases. But, this is difficult because the considered sequence \( \{ \hat{x}_t, A^i_t (\hat{x}_t) \} \) is different from the actual sequence of context and selected arms \( \{ x_t, a^i_t \} \) under MinWD. To circumvent this issue, we first introduce the concept of e-covering (Wu2016). Denote \( \Phi = \mathcal{X} \times \mathcal{A} \) as the context-arm space. If a finite set \( \Phi_e \) is an e-covering of the space \( \Phi \), then for each \( \varphi \in \Phi \), there exists at least one \( \varphi_e \in \Phi_e \) satisfying \( \| \varphi - \varphi_e \|_2 \leq e \). Define \( C_e (\varphi_e) = \{ \varphi \in \Phi | \| \varphi - \varphi_e \|_2 \leq e \} \) as the cell with respect to \( \varphi_e \). Since the dimension of the entries in \( \Phi \) is \( d \), the size of the \( \Phi_e \) is \( | \Phi_e | \sim O \left( \frac{\log d}{e^2} \right) \). Besides, we assume the mapping function \( \phi \) 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 \( \tilde{\mathcal{X}}_T = \{ \tilde{x}_{a_1,1}, \cdots, \tilde{x}_{a_T,T} \} \) be the considered sequence of contexts and actions. Suppose that both \( x_{a_t,t} \) and \( \tilde{x}_{a_t,t} \) belong to \( \Phi \). Besides, with an e-covering \( \Phi_e \subseteq \Phi \), \( e > 0 \), there exists \( \kappa > 0 \) such that two conditions are satisfied: First, \( \forall \varphi \in \Phi_e, \exists \varphi_e \in \Phi_e \) such that \( | \varphi - \varphi_e | \leq \frac{\kappa}{e} \). Second, if at round \( t \), \( x_{a_t,t} \in C_e (\varphi) \) for some \( \varphi \in \Phi_e \), then \( \exists \varphi_e \in C_e (\varphi) \) such that \( x_{a_t,t} \in C_e (\varphi_e) \). If the mapping function \( \phi \) is Lipschitz continuous with constant \( L_{\phi} \), the sum of squared confidence widths is bounded as

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

where \( d \) is the dimension of \( x_{a_t,t}, \tilde{d} \) is the effective dimension defined in the proof, \( s_t^2 (\hat{x}_{a_t,t}, \tilde{a}_t) = \phi(\hat{x}_{a_t,t}, \tilde{a}_t)^T V_{t-1} \phi(\hat{x}_{a_t,t}, \tilde{a}_t) \) and \( V_t = \lambda I + \sum_{s=1}^{t} \phi(x_{a_t,s}) \phi(x_{a_t,s})^T \).

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 \( \left[ \kappa/e^2 \right] \), which is proportional to the size of the e-covering \( | \Phi_e | \). 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 \( A^i_t (\hat{x}_t) \) is the UCB-optimal arm, which becomes close to the optimal arm for \( \hat{x}_t \) 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 \( x_t \in B_{\kappa} (\hat{x}_t) \) 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:

\[
R_T \leq \sum_{t=1}^{T} M R_t + 2h_t T^2 \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{1}{\lambda} \right) + \frac{4\sqrt{2\lambda} L_{\phi} \kappa^2/h_t T^{1-1/2}}{d} + 2d \log \left( 1 + \frac{T}{d\lambda} \right),
\]

where \( M R_t \) 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, \( \tilde{d} \) is the rank of \( K_t \), and \( h_t \) is given in Lemma 4.

Remark 4. Theorem 6 shows that by MinWD, \( R_T = \sum_{t=1}^{T} M R_t \) is sub-linear w.r.t. \( T \) and Type-II robustness is effectively achieved according to Definition 4. This means the true regret bound approaches \( \sum_{t=1}^{T} M R_t \), 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 \( x_t \in B_{\kappa} (\hat{x}_t) \) 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 cumulative reward

\[
F_T \geq \sum_{t=1}^{T} \left[ M R_t - M R_t \right] - 2h_t T^2 \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{1}{\lambda} \right) - \frac{4\sqrt{2\lambda} L_{\phi} \kappa^2/h_t T^{1-1/2}}{d} - 2h_t T^2 \log \left( 1 + \frac{T}{d\lambda} \right)
\]

where \( M R_t \) 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, \( d \) is the rank of \( K_t \), and \( h_t \) is given in Lemma 4.

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 \( M R_t \). With perfect context, we have \( M R_t = 0 \), and hence MaxMinUCB and MinWD both asymptotically maximize the reward, implying that these two types of robustness are the same under perfect context.

5.3 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...
Table 1: Summary of Analysis

| Algorithms   | Regret                           | Reward                          |
|--------------|----------------------------------|---------------------------------|
| MaxMinUCB    | $\sum_{t=1}^{T} MR_t + O(\sqrt{T \log T})$ | $\sum_{t=1}^{T} MF_t - O(\sqrt{T \log T})$ |
| MinWD        | $\sum_{t=1}^{T} MR_t + O(T^{1/4} \log T + T^{-1/2} + \sqrt{T \log T})$ | $\sum_{t=1}^{T} [MF_t - MR_t] + O(T^{1/4} \sqrt{T \log T} + T^{-1/2} + \sqrt{T \log T})$ |

Figure 2: Different cumulative regret objectives for different algorithms.

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). This is because the regret of MinWD approaches the optimal worst-case regret (Theorem 4). 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.

6 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 $\mu_a$, the computation latency satisfies an $\text{M}/\text{M}/1$ queueing model and the average communication delay between this datacenter and users is $\rho_a$. 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. 4, worst-case regret in Eqn. 6 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 $B_\Delta(x_t)$ at each round. In the simulations, Gaussian kernel with parameter 0.1 is used for reward (loss) estimation. $\lambda$ in Eqn. 3 is set as 0.1. In Fig. 2(b), we compare different algorithms in terms of three cumulative regret objectives: robust regret in Eqn. 4, worst-case regret in Eqn. 6 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 $B_\Delta(x_t)$ at each round. In the simulations, Gaussian kernel with parameter 0.1 is used for reward (loss) estimation. $\lambda$ in Eqn. 3 is set as 0.1. In Fig. 2(c), we compare different algorithms in terms of three cumulative regret objectives: robust regret in Eqn. 4, worst-case regret in Eqn. 6 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 $B_\Delta(x_t)$ at each round. In the simulations, Gaussian kernel with parameter 0.1 is used for reward (loss) estimation. $\lambda$ in Eqn. 3 is set as 0.1.

# 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.

Amazon. 2021. Amazon AWS Auto Scaling Documentation. [https://docs.aws.amazon.com/autoscaling/](https://docs.aws.amazon.com/autoscaling/)

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*.

Garcia, 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.

Guo, 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*. 
A key novelty of our work is the consideration of imperfect context for arm selection, which characterizes many practical applications such as resource management problems where the true context is difficult to obtain for arm selection until revealed later. We list some examples of these applications in this section and provide the simulation settings.

A.1 Motivating Applications

Cloud Resource Management. Cloud computing platforms are crucial infrastructures offering utility-style computation resources to users on demand. To optimize the performance metrics such as latency and cost for these applications, efficient online cloud resource management such as dynamic virtual machine scheduling is necessary. Contextual bandit learning can be employed in this scenario where the exact workload information (measured in, e.g., how many requests will arrive per unit time) is crucial context information, but cannot be known until the workload actually arrives. Instead, the agent can only predict the upcoming workload by exploiting the recent workload history plus other applicable system features. A real-world example is Amazon’s predictive scaling that leverages time series prediction to estimate upcoming workload for virtual machine scheduling. In this motivating example, the context prediction error comes primarily from the workload predictor.

Energy Scheduling in Smart Grid. Energy load is a crucial context information for energy scheduling in smart grid. However, 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 (i.e., context for energy scheduling) with higher-than-normal errors. Thus, our model also captures a novel adversarial setting where erroneously predicted context is presented to the agent for arm selection. This example shows that the use of machine learning-based predictor for acquiring context to facilitate the agent’s arm selection can also open a new attack surface — an outside attacker may adversarially modify the predicted context — which requires a robust algorithm with a provable worst-case performance guarantee.

Online Edge Datacenter Selection. With the rapidly increasing number of devices at the Internet edge, computational demand by latency-sensitive applications (e.g., assisted driving and virtual reality) has been escalating. Edge computing is a promising solution to meet the demand, which deploys computation resources at densely-distributed edge datacenters close to end users and thus reduces the overall latency (Shi et al. 2016). Since users’ workloads can be processed in multiple edge datacenters, dynamic selection of edge datacenters plays a key role for minimizing the overall latency — given multiple heterogeneous edge datacenters located in different locations, which one should be selected? Here, servers in each edge datacenter can be either virtual machines rented from third-party service providers or physical servers owned by the edge computing provider itself. We refer to this problem as online edge datacenter selection. Our considered bandit setting applies to the problem of online edge datacenter selection. 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 dynamically select edge datacenters to optimize the latency in a robust manner given imperfect information about users’ workloads.
A.2 Simulation Settings in Section 6

In our simulation, we apply our algorithms to online edge datacenter selection in the context of edge computing. We consider users’ workloads can be processed in one of four available edge datacenters, each having different computation capabilities and communication latencies. Here, we consider a simple latency model to capture first-order effects. Concretely, we assume that the service rate of the edge datacenter $a$, $a \in \mathcal{A} = \{1, 2, 3, 4\}$, is $\mu_a$, the computation latency satisfies an M/M/1 queueing model and the average communication delay between this datacenter and users is $p_a$. The values of $\mu_a$ and $p_a$ are shown in Table 2. In the simulations, perfect context sequence $\{x_t\}$ is generated by sampling i.i.d. uniform distribution between 10 and 30 for each round, while the defense budget $\Delta$ is set as 2. The average total latency cost can be expressed as

$$f(x, a) = \frac{x}{\mu_a - x} + p_a \cdot x$$  \hspace{1cm} (18)$$

whose inverse can be equivalently viewed as the reward in our model.

| Datacenter $a$ | I | II | III | IV |
|---------------|---|----|-----|----|
| $\mu_a$       | 35 | 38 | 45  | 51 |
| $p_a$         | 0.04 | 0.05 | 0.07/4 | 0.088 |

While we use the simple latency model in Eqn. (18) for generating the ground-truth cost, the function form and parameters (e.g., $\mu_a$) may not be known to the agent and needs to be learnt based on latency feedback and revealed context. Note that some practical factors (e.g., workload parallelism) are beyond the scope of our analysis and incorporating them into our simulation does not add substantially to our main contribution.

B Algorithm and Proofs Related to SimpleUCB

B.1 Algorithm of SimpleUCB

We describe SimpleUCB in Algorithm 2.

Algorithm 2 Simple UCB (SimpleUCB)

\begin{algorithm}
\begin{algorithmic}
\FOR{$t = 1, \ldots, T$}
\STATE Receive imperfect context $\hat{x}_t$.
\STATE Select arm $a^*_t$ as the UCB-optimal arm $A^*_t(\hat{x}_t)$ defined in Definition 2.
\STATE Observe the true context $x_t$ and the reward $y_t$.
\ENDFOR
\end{algorithmic}
\end{algorithm}

B.2 Proof of Lemma 1

Lemma 1 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 $h_t = \sqrt{\lambda B + b/\gamma_t - 2 \log(\delta)}$, $\gamma_t = \log \det(I + K_t/\lambda) \leq d \log(1 + d/\lambda)$ and $d$ is the rank of $K_t$.

Let $V_t = \lambda I + \sum_{s=1}^t \phi(x, a) \phi(x, a)^\top$, the squared confidence width is given by $\phi(x, a) = \phi(x, a)^\top V_t^{-1} \phi(x, a) = 1/k([x, a], [x, a]) - 1/k_t(x, a)^\top (K_t + \lambda I)^{-1} k_t(x, a)$.

Proof. Let $\phi : (\mathcal{X} \times \mathcal{A}) \to \mathcal{H}$ be the mapping function with respect to $k(\cdot, \cdot)$. Define $\Psi_t$ with the sth row as $\phi(x_s, a_s), s = 1, 2, \ldots, t - 1$ and thus $K_t = \Psi_t^\top \Psi_t$. Denote $y_t = [y_1, \ldots, y_{t-1}]^\top$ as the collected rewards and the noise vector $n_t = [n_1, n_2, \ldots, n_{t-1}]^\top$, so $y_t = \Psi_t \theta + n_t$.

Then we have

$$|f(x, a) - \hat{f}_t(x, a)| = |f(x, a) - k_t^\top(x, a)(K_t + \lambda I)^{-1} y_t|$$

$$= |f(x, a) - k_t^\top(x, a)(K_t + \lambda I)^{-1}(\Psi_t \theta + n_t)|$$

$$\leq |f(x, a) - k_t^\top(x, a)(K_t + \lambda I)^{-1}\Psi_t \theta| +$$

$$|k_t^\top(x, a)(K_t + \lambda I)^{-1} n_t|,$$

(19)

where $I_t$ is an identity matrix with $(t - 1)$ dimensions and the last inequality comes from triangle inequality.

Let $V_{t-1} = \Psi_t^\top \Psi_t + \lambda I$ and define $\phi_t(x, a) = \phi(x, a)^\top V_{t-1}^{-1} \phi(x, a)$. By Woodbury formula, we can write $\phi_t(x, a)$ by kernel functions:

$$s^2_t(x, a) = \phi(x, a)^\top \left(1 - \frac{1}{\lambda} \frac{1}{\Psi_t^\top (\lambda I + \Psi_t \Psi_t^\top)^{-1} \Psi_t}\right) \phi(x, a)$$

$$= 1/\lambda k([x, a], [x, a]) - 1/\lambda k_t(x, a)^\top (K_t + \lambda I)^{-1} k_t(x, a).$$

(20)

For the first term of Eqn. (19), using Woodbury formula, we have

$$|f(x, a) - k_t^\top(x, a)(K_t + \lambda I)^{-1} \Psi_t \theta|$$

$$= |\phi(x, a)^\top \theta - \phi(x, a)^\top \Psi_t \Psi_t^\top (\Psi_t \Psi_t^\top + \lambda I)^{-1} \Psi_t \theta|$$

$$= |\phi(x, a)^\top \theta - \phi(x, a)^\top (\Psi_t \Psi_t^\top + \lambda I)^{-1} \Psi_t \theta|$$

$$= |\lambda \phi(x, a)^\top (\Psi_t \Psi_t^\top + \lambda I)^{-1} \theta| \leq \sqrt{\lambda} B s_t(x, a),$$

(21)

where the inequality comes from Cauchy-Schwartz inequality and $\|\theta\|_{V_{t-1}}^{-1} \leq \|\theta\|_{V_{t-1}} \leq \frac{\beta}{\lambda}$.

For the second term, we have the following inequalities.

$$|k_t^\top(x, a)(K_t + \lambda I)^{-1} n_t|$$

$$\leq \|\phi(x, a)\|_{V_{t-1}^{-1}} \sqrt{n_t^\top \Psi_t \Psi_t^\top \Psi_t + \lambda I)^{-1} \Psi_t n_t}$$

$$= s_t(x, a) \|n_t\|_{K_t(K_t + \lambda I)^{-1}},$$

(22)

where the first inequality comes from Cauchy-Schwartz inequality.
Since \( n_t \) satisfies \( b \)-sub-Gaussian distribution conditioned on \( F_{t−1} \), by Theorem 1 in (Abbasi-Yadkori, Pál, and Szepesvári[2011]), with probability \( 1−\delta, \delta \in (0, 1) \) for \( t \in \mathbb{N} \) and \( a \in A \), we have
\[
\|n_t\|_F^2 |K_t(K_t+\lambda I_t)^{-1} - \|V_t^{-1}\|^2_{F} \\
\leq 2b^2 \log \left( \frac{\det(V_t^{-1})^{1/2}}{\delta \det(\lambda I_t)^{1/2}} \right) = b^2 \left( \gamma_t - 2 \log(\delta) \right).
\] (23)
where \( \gamma_t = \log \det \left( K_t/\lambda + I_t \right) \leq \hat{d} \log(1 + \frac{T}{d}) \), and \( \hat{d} \) is the dimension of \( K_t \), the inequality comes from Lemma 10 in [Abbasi-Yadkori, Pál, and Szepesvári[2011]].

Combining the bounds of the first and second term in Eqn. (19) and let \( h_t = \sqrt{d} + b \sqrt{\gamma_t - 2 \log(\delta)} \), we obtain the concentration bound in Lemma 3.1: \( |f(x, a) - \hat{f}_t(x, a)| \leq h_t s_t(x, a) \).

B.3 Proof of Lemma 2

Lemma 2 The 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 \( \gamma_T = \log \det \left( I_T + K_T/\lambda \right) \). By Hölder's inequality, we have
\[
\sum_{t=1}^{T} s_t^2(x_t, a_t) \leq \left( \sum_{t=1}^{T} s_t^2(x_t, a_t) \right)^{1/2} \left( T \sum_{t=1}^{T} s_t^2(x_t, a_t) \right)^{1/2} = \sqrt{2T} \gamma_T.
\]
Proof. By Lemma 11 in [Abbasi-Yadkori, Pál, and Szepesvári[2011]], if \( \lambda \geq 1 \), we have
\[
\sum_{t=1}^{T} s_t^2(x_t, a_t) = \sum_{t=1}^{T} \phi(x_t, a_t)^T V_t^{-1} \phi(x_t, a_t)
\leq 2 \log \det \left( \frac{1}{\lambda} V_t \right) = 2\gamma_T
\]
where \( \gamma_T = \log \det \left( K_T/\lambda + I_T \right) \). By Hölder's inequality, we have
\[
\sum_{t=1}^{T} s_t(x_t, a_t) \leq \sqrt{T} \sum_{t=1}^{T} s_t^2(x_t, a_t) \leq \sqrt{2T} \gamma_T.
\]

C Proofs Related to MaxMinUCB

C.1 Proof of Theorem 3

Theorem 3 If MaxMinUCB is used to select arms with imperfect context, then for any true contexts \( x_t \in B_\Delta(\hat{x}_t) \) 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
\[
F_T \geq \sum_{t=1}^{T} M F_t - 2h_t \sqrt{2T \hat{d}} \log(1 + \frac{T}{d})
\]
where \( M F_t \) is the optimal worst-case reward in Eqn. (6), \( \hat{d} \) is the rank of \( K_t \), and \( h_t \) is given in Lemma 2.
Proof. By Lemma 1 and UCB defined in Definition 1, with a probability \( 1−\delta, \delta \in (0, 1) \), the instantaneous regret can be bounded as
\[
f(x_t, a_t) \geq U_t(x_t, a_t) - 2h_t s_t(x_t, a_t)
\geq U_t(x_t, a_t) - 2h_t s_t(x_t, a_t)
\]
where the last inequality comes from the arm selection policy of MaxMinUCB implying that \( x_t = \arg \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a_t) \).

With the definition of the optimal worst-case reward \( M F_t \), and exploiting the arm selection policy of MaxMinUCB, we can further bound the instantaneous reward as below.
\[
f(x_t, a_t) \geq \left[ U_t(x_t, a_t) - 2h_t s_t(x_t, a_t) \right]
\geq \left[ U_t(x_t, a_t) - 2h_t s_t(x_t, a_t) \right]
\geq \left[ U_t(x_t, a_t) - 2h_t s_t(x_t, a_t) \right] - f(x_t, a_t) + M F_t
\geq M F_t - 2h_t s_t(x_t, a_t)
\]
where \( a_t = \arg \max_{x \in A} \min_{x \in B_\Delta(\hat{x}_t)} f(x, a) \) is the optimal arm for maximizing the worst-case reward defined in Definition 1, \( x_t = \arg \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a_t) \), the second inequality comes from the arm selection strategy of MaxMinUCB such that \( \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a_t) \geq \min_{x \in B_\Delta(\hat{x}_t)} U_t(x, a_t) \), and the third inequality holds because the definition of \( M F_t \) in Eq. (6), which guarantees \( M F_t = \min_{x \in B_\Delta(\hat{x}_t)} f(x, a_t) \leq f(x_t, a_t) \), and the last inequality comes from Lemma 1 which guarantees \( U_t(x_t, a_t) \geq f(x_t, a_t) \). Therefore, combined with Lemma 2, we can get the bound on the cumulative reward of MaxMinUCB.

D Proofs Related to MinWD

D.1 Proof of Lemma 4

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
\[
\mathcal{D}_{a_t} \leq M R_t + 2h_t s_t(x_t, a_t)
\]
where \( \mathcal{D}_{a_t} \) is the ability of the selection policy of MinWD to select arms with imperfect context.
where $MR_t$ is the optimal worst-case regret defined in Eqn. (11), $\hat{x}_t = \arg\max_{x \in B_t} D_{\hat{a}_t}(x)$ is the context that maximizes the degradation given the arm $\hat{a}_t$ defined for the optimal worst-case regret in Eqn. (10).

**Proof.** Recall that in Eqn. (9), the optimal arm for minimizing the worst-case regret is $\hat{a}_t = \arg\min_{a \in A} \max_{x \in B_t} r(x, a)$. By the arm selection policy of MinWD, $a^o_{t, t}$ is the arm that minimizes the degradation given the arm $\hat{a}_t$ for the worst-case regret defined in Eqn. (11) and the second inequality is from Lemma 1 which guarantees that $\exists$ such that $
abla$.

The set $\Phi$ is a finite set. Furthermore, we assume that $\Phi$ is uniformly distributed. Thus,

$$\gamma(W_{l,m}, \phi) = \log \frac{1}{\lambda} \left( \frac{M}{\lambda} \right),$$

for $\gamma(W_{l,m}, \phi)$ defined in Eqn. (30), where $M$ is the number of groups and $\lambda$ is the maximum number of arms in each group.

**Proof.** Let $M = \lceil T / |\Phi| \rceil$. Divide the round stamp sequence $\{1, 2, \ldots, M \}$ uniformly into $|\Phi|$ groups, each with $M$ elements. The $i$th group of round stamps, $1 \leq i \leq |\Phi|$, is $G_i = \{l, l + |\Phi|, \ldots, l + (M - 1)|\Phi|\}$, and $G_{i,m} = \{l, l + |\Phi|, \ldots, l + (m - 1)|\Phi|\}$, $1 \leq m \leq M$ is a subset of $G_i$ with the first $m$ entries. A simple example of group construction is shown in Fig. 3.

Let $W_{l,m} = \gamma(W_{l,m}, G_i)$.

Let’s consider the sequence in $G_i$. The two conditions in Lemma 8 imply that for $t \in G_i$ and its index $m$, $i$, $s \in G_{i,m}$, if $\hat{x}_{a,s} = \phi$, there exists $s' = s + |\Phi|$, such that the true context-arm $x_{a,s'} = \phi$. Therefore we have $s' \in T_{i-1} = \{1, 2, \ldots, T - 1\}$, and we conclude that $x_{a,s'} \in G_{i,m-1} = \{1, 2, \ldots, |\Phi|\}$. Thus, considering $V_{t-1}$ and $W_{l,m-1}$ are both positive definite matrices, we have $V_{t-1} \succeq W_{l,m-1}$ and $W_{l,m-1} \succeq V_{t-1}$. Therefore, for group $i (1 \leq i \leq |\Phi|)$, we have

$$\sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \gamma(W_{l,m}, \phi) \leq \sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \gamma(W_{l,m}, \phi)$$

From the above analysis, since $s_t^2(\hat{x}_{a,t}) \leq 1/\lambda$, we can get

$$\sum_{t=1}^T s_t^2(\hat{x}_{a,t}) = \frac{R}{|\Phi|} \sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \frac{R}{|\Phi|} \sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \gamma(W_{l,m}, \phi) \leq \sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \gamma(W_{l,m}, \phi) \leq \sum_{t \in G_i} s_t^2(\hat{x}_{a,t}) \leq \gamma(W_{l,m}, \phi) \leq \sum_{t \in G_i} s_t^2(\hat{x}_{a,t})$$

where $s_t^2(\hat{x}_{a,t}) = \hat{x}_{a,t}^T V_{t-1}^{-1} \hat{x}_{a,t}$ and $V_t = \lambda I + \sum_{s=1}^{M} x_{a,s} x_{a,s}^T$. Figure 3: An example of sequence grouping in Lemma 8. The feature set is $\Phi = \{z_1, z_2\}$, $\kappa = 2$, $|\Phi| = 4$ groups are constructed: $G_1 = \{1, 5, 9, \ldots\}$, $G_2 = \{2, 6, 10, \ldots\}$, $G_3 = \{3, 7, 11, \ldots\}$, $G_4 = \{4, 8, \ldots\}$. The interval between the features in considered sequence and their counterparts in true sequence is less than $|\Phi|$.
where the second inequality comes from Lemma 11 in (Abbasi-Yadkori, Pál, and Szepesvári [2011]), thus completing the proof. □

If the context-arm space is continuous, the finite set assumption of \( \Phi \) in Lemma 7 is not satisfied anymore. To overcome this issue, we construct an \( \epsilon \)-covering \( \Phi_\epsilon \subset \Phi \) for the context-arm space \( \Phi \) such that for each \( \varphi \in \Phi \), there exists at least one \( \bar{\varphi} \in \Phi_\epsilon \) satisfying \( \| \varphi - \bar{\varphi} \| \leq \epsilon \). Since the dimension of \( x_{a,t} \) is \( d \), the size of the \( \epsilon \)-covering is \( |\Phi_\epsilon| \sim O \left( \frac{1}{\epsilon^d} \right) \). Now we can bound the sum of confidence width with linear reward function and continuous context-arm space in Lemma 8.

**Lemma 8 (Sum of Confidence Width with Continuous Context-Arm Space).** Let \( X_T = \{ x_{a_1, \cdot} , \cdots , x_{a_T,T} \} \) be the sequence of true contexts and selected arms by bandit algorithms and \( \bar{X}_T = \{ \bar{x}_{a_1,\cdot} , \cdots , \bar{x}_{a_T,T} \} \) be the considered sequence of contexts and actions. Suppose that both \( x_{a,t} \) and \( \bar{x}_{a,t} \) belong to \( \Phi \). Besides, with an \( \epsilon \)-covering \( \Phi_\epsilon \subset \Phi \), \( \epsilon > 0 \), there exists \( \kappa \geq 0 \) such that two conditions are satisfied for \( \bar{X}_T \). First, \( \forall \varphi \in \Phi_\epsilon \), \( x_{a,t} \leq \left\lceil \frac{\kappa}{\epsilon d} \right\rceil \) such that \( x_{a,t} \in C_{\bar{\varphi}}(\bar{\varphi}) \). Second, if at round \( t \), \( x_{a,t} \in C_{\bar{\varphi}}(\varphi) \) for some \( \varphi \in \Phi_\epsilon \), then \( \exists \bar{\varphi} \in \Phi_\epsilon \) such that \( x_{a,t} \in C_{\bar{\varphi}}(\bar{\varphi}) \). The sum of squared confidence width is bounded as

\[
\sum_{t=1}^{T} s_{t}^{2}(\hat{x}_{a,t}) \leq \sqrt{T} \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{8\epsilon^{2d}/\lambda}{T^{1-1/d}} \right),
\]

where \( s_{t}^{2}(\hat{x}_{a,t}) = \sum_{s=1}^{L_{d}} x_{a,s}x_{a,s}^{T} \) and \( \Phi = \left\{ x_{a_1, \cdot} , \cdots , x_{a_T,T} \right\} \)

Proof: Based on the \( \epsilon \)-covering \( \Phi_\epsilon \subset \Phi \), we divide the round stamp sequence \( \{1,2,M \left\lceil \kappa/\epsilon d \right\rceil \} \) uniformly into \( \left\lceil \kappa/\epsilon d \right\rceil \) groups, each with \( M = \left[ T/ \left\lceil \kappa/\epsilon d \right\rceil \right] \) elements. The \( l \)-th group \( \{1 \leq l \leq \left\lceil \kappa/\epsilon d \right\rceil \} \) is \( \Gamma_{l} = \{1, l + \left\lceil \kappa/\epsilon d \right\rceil, \ldots , l + (M - 1) \left\lceil \kappa/\epsilon d \right\rceil \} \), and \( \Gamma_{l,m} = \{1, l + \left\lceil \kappa/\epsilon d \right\rceil, \ldots , l + (m - 1) \left\lceil \kappa/\epsilon d \right\rceil \} \) is a subset of \( \Gamma_{l} \) with the first \( m \) entries.

Now we consider the \( l \)-th group. The two conditions in this lemma imply that for a certain \( t \in \Gamma_{l} \) and its index \( m_{t} = \left\lceil (t-1)/ \left\lceil \kappa/\epsilon d \right\rceil \right\rceil \), \( \forall s \in \Gamma_{l,m_{t}-1} \), \( \exists \hat{x}_{a,s} \in C_{\bar{\varphi}}(\bar{\varphi}) \) for some \( \varphi \in \Phi_\epsilon \), there exists \( s' \) such that \( s' \leq s < \left\lceil \kappa/\epsilon d \right\rceil \) such that the true context-arm \( x_{a,s}, x_{a,s'} \in C_{\bar{\varphi}}(\bar{\varphi}) \). Denote \( \zeta(s) = \zeta(s) \) mapping \( s \) to its corresponding \( s' \), and we have \( \zeta(s) \leq t - 1 \). Also, denote \( \Gamma_{l,m_{t}-1} = \{ \zeta(s) \mid s \in \Gamma_{l,m_{t}-1} \} \), and we have \( \Gamma_{l,m_{t}-1} \subseteq \hat{T}_{m} = \{1, 2, \ldots , t - 1\} \). Let \( \Phi_{l,m_{t}} = \lambda_{d}d + \sum_{s' \in \Gamma_{l,m_{t}} \times \nabla_{c}} x_{a,s',c}x_{a,s',c}^{T} \), for \( 1 \leq l \leq \left\lceil \kappa/\epsilon d \right\rceil \). By the above analysis, we have \( \sum_{t \in \Gamma_{l,m_{t}-1}} \leq \sum_{t \in \Gamma_{l,m_{t}-1}} \leq \nabla_{c}^{2} \nabla_{c}^{T} \), and so \( \sum_{t \in \Gamma_{l,m_{t}-1}} \leq \nabla_{c}^{2} \nabla_{c}^{T} \), considering \( \sum_{t \in \Gamma_{l,m_{t}-1}} \) and \( \sum_{t \in \Gamma_{l,m_{t}-1}} \) are both positive definite matrices.

By Lemma 2, for the sequence \( \{x_{a_{t}(c),t}(t) \in \Gamma_{l} \} \), we have \( \sum_{t \in \Gamma_{l}} x_{a_{t}(c),t}(t) \leq \gamma(\nabla_{c}^{2}) \).

where the last inequality holds because \( \gamma(\nabla_{c}^{2}) \leq 2d \log \left( 1 + \frac{T}{d\lambda} \right) \) by Lemma 11 in (Abbasi-Yadkori, Pál, and Szepesvári [2011]), and \( M = \left[ T/ \left\lceil \kappa/\epsilon d \right\rceil \right] \). Let \( \kappa/\epsilon d \leq T^{1/2} \), i.e. \( \epsilon \leq \left( \kappa/2T \right)^{1/2d} \), then we have

\[
\sum_{t=1}^{T} s_{t}^{2}(\hat{x}_{a,t}) \leq \sqrt{T} \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{8\epsilon^{2d}/\lambda}{T^{1-1/d}} \right),
\]

thus completing the proof. □

Now we can prove proposition 5 by generalizing Lemma 8 into kernel case assuming the mapping function \( \phi : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{H} \) is Lipschitz continuous with constant \( L_{d} \), i.e. \( \forall x,y \in \mathcal{X} \times \mathcal{A} \), \( \| \phi(x) - \phi(y) \| \leq L_{d} \| x - y \| \).

**Proof of Proposition 5.**

**Proposition 5.** Let \( X_T = \{ x_{a_1, \cdot} , \cdots , x_{a_T,T} \} \) be the sequence of true contexts and selected arms by bandit algorithms and \( \bar{X}_T = \{ \bar{x}_{a_1,\cdot} , \cdots , \bar{x}_{a_T,T} \} \) be the considered sequence of contexts and actions. Suppose that both \( x_{a,t} \) and \( \bar{x}_{a,t} \) belong to \( \Phi \). Besides, with an \( \epsilon \)-covering \( \Phi_\epsilon \subset \Phi \), \( \epsilon > 0 \), there exists \( \kappa \geq 0 \) such that two conditions are satisfied. First, \( \forall \varphi \in \Phi_\epsilon \), \( \exists \bar{\varphi} \in \Phi_\epsilon \) such that \( x_{a,t} \in C_{\bar{\varphi}}(\bar{\varphi}) \).
Second, if at round $t$, $x_{a,t} \in C_{\bar{\varphi}}(\bar{\varphi})$ for some $\bar{\varphi} \in \Phi$, then
\[ \exists t' < t + [\kappa/\epsilon^4]\] such that $x_{a,t'} \in C_{\bar{\varphi}}(\bar{\varphi})$. If the mapping function $\hat{\varphi}$ is Lipschitz continuous with constant $L_\phi$, the sum of squared confidence widths is bounded as
\begin{equation}
\sum_{t=1}^{T} s_t^2(\hat{x}_{a,t}) \leq \sqrt{T} \left( 4d \log \left( 1 + \frac{T}{\lambda} \right) + \frac{1}{\lambda} \right) + \frac{8L_\phi^2 \kappa^{2/d}}{\lambda} T^{-1/d},
\end{equation}
where $d$ is the dimension of $x_{a,t}$, $\tilde{d}$ is the effective dimension defined in the proof, $s_t^2(\hat{x}_{a,t}) = \phi(\hat{x}_{a,t})^T V_{t-1}^{-1} \phi(\hat{x}_{a,t})$ and $V_t = \lambda I + \sum_{s=1}^{d} \phi(x_{a,s}) \phi(x_{a,s})^T$.

**Proof.** With the same method in Lemma 8, we construct $\left[ \kappa/\epsilon^4 \right]$ groups, each with $M = \left[ T/\left[ \kappa/\epsilon^4 \right] \right]$ elements. Define $G_t, G_t^0, G_t^0 = (t, m, \epsilon_t)$ same as Lemma 8 and let $W_{t,m} = \lambda I + \sum_{s=1}^{d} \phi(x_{a,s}) \phi(x_{a,s})^T = \lambda I + \sum_{s=1}^{d} \phi(x_{a,s}) \phi(x_{a,s})^T$, for $1 \leq l \leq [\kappa/\epsilon^4]$.

By Lemma 8 we have $V_{t-1} \geq W_{t,m_{t-1}}$, and so $V_{t-1} \geq W_{t,m_{t-1}}^{-1}$. Therefore for group $G_t$, we have
\begin{align*}
\sum_{t=1}^{T} s_t^2(\hat{x}_{a,t}) &= \sum_{t \in G_t} \phi(\hat{x}_{a,t})^T V_{t-1}^{-1} \phi(\hat{x}_{a,t}) \\
&= \sum_{t \in G_t} \| \phi(\hat{x}_{a(t)}, \zeta(t)) + \phi(\hat{x}_{a,t}) - \phi(x_{a(t)}, \zeta(t)) \|^2_{V_{t-1}^{-1}} \\
&\leq 2 \sum_{t \in G_t} \| \phi(\hat{x}_{a(t)}, \zeta(t)) \|^2_{V_{t-1}^{-1}} + 2 \sum_{t \in G_t} \| \phi(\hat{x}_{a,t}) - \phi(x_{a(t)}, \zeta(t)) \|^2_{V_{t-1}^{-1}} \\
&\leq 2 \sum_{t \in G_t} \| \phi(x_{a(t)}, \zeta(t)) \|^2_{W_{t,m_{t-1}}} + 2 \frac{M}{\lambda} \left( L_\phi \| \epsilon_t \|_2 \right)^2 \\
&\leq 2 \lambda \| \epsilon_t \|_2^2 + \frac{8M^2 \kappa^{2/d}}{\lambda} T^{-1/d},
\end{align*}
where the first inequality comes from Cauchy-Schwartz inequality and the second inequality holds because $V_{t-1}^{-1} \geq W_{t,m_{t-1}}^{-1}$, eig($V_{t-1}^{-1}$) $\leq \frac{1}{\lambda}$, and Lipschitz continuity of $\phi$ such that $\| \phi(\hat{x}_{a,t}) - \phi(x_{a(t)}, \zeta(t)) \|_2 \leq L_\phi \| \epsilon_t \|_2$, and the last inequality holds by Lemma 2.

For kernel case, with $K_{t, M} \in R^{M \times M}$ containing $k(x_{a(t)}, \zeta(t), x_{a(t)}, \zeta(t))$ for $t \in G_t$, we have $\gamma(W_{t,M}) = \log \frac{\det(W_{t,M})}{\det(M)} = \log \frac{\det(K_{t, M})}{\det(M)} \leq 2d \log (1 + \frac{M}{d})$, where $d$ is the rank of $K_{t, M}$. Define the effective dimension as $\tilde{d} = \arg \max_{d} (d, \cdots, d \in [a/d])$, then the kernel case uses the sum of squared confidence widths for the whole sequence as
\begin{equation}
\sum_{t=1}^{T} s_t^2(\hat{x}_{a,t}) = \sum_{t=1}^{T} \sum_{\tilde{t} \in G_t} s_{\tilde{t}}^2(\hat{x}_{a,\tilde{t}}) + \sum_{t=1}^{T} s_t^2(\hat{x}_{a,t}) \\
\leq [\kappa/\epsilon^4] \left( 2 \lambda \| \epsilon_t \|_2^2 + \frac{1}{\lambda} \right) \\
\leq [\kappa/\epsilon^4] \left( 4d \log (1 + \frac{M}{d}) + \frac{1}{\lambda} \right) + \frac{8M^2 \kappa^{2/d}}{\lambda} T^{-1/d},
\end{equation}
thus completing the proof. \(\square\)

**D.3 Proof of Theorem 6**

**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:
\begin{equation}
R_T \leq 4 \sqrt{\frac{d}{\lambda} \| \epsilon_t \|^2_{W_{t,M}}} + 2 \sqrt{2T \epsilon^2},
\end{equation}
where $MR_t$ is the optimal worst-case regret for round $t$ in Eqn. (11), $d$ is the dimension of $x_{a,t}$, $\tilde{d}$ is the effective dimension defined in the proof of Proposition 5, $\tilde{d}$ is the rank of $K_t$, and $h_T$ is given in Lemma 7.

**Proof.** Since with a probability $1 - \delta$, $\delta \in (0, 1)$, $r(x_t, a_{t}^0) \leq D_{a_{t}^0} + 2h_t s_t(x_t, a_{t}^0)$, we can bound the cumulative regret of MinWD as
\begin{equation}
R_T = \sum_{t=1}^{T} r(x_t, a_{t}^0) \leq \sum_{t=1}^{T} \left[ D_{a_{t}^0} + 2h_t s_t(x_t, a_{t}^0) \right] \\
\leq \sum_{t=1}^{T} \left[ MR_t + 2h_t \hat{s}_t(\hat{x}_t, a_{t}^0) \right],
\end{equation}
where the second inequality comes from Lemma 4.

Since the sequence $\{x_t, a_{t}^0\}$ meets the conditions in Proposition 5, by replacing $\hat{x}_{a,t}$ and $x_{a,t}$ in Proposition 5,
with \([\hat{x}_t, A^t(\hat{x}_t)]\) and \([x_t, a^t]\), we have
\[
\sum_{t=1}^{T} s_t \left( \hat{x}_t, A^t(\hat{x}_t) \right) \leq \left( T \sum_{t=1}^{T} s^2_t \left( \hat{x}_t, A^t(\hat{x}_t) \right) \right)^{1/2} \leq \sqrt{T^{3/2} \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{1}{\lambda} \right) + \frac{8L_{2\phi}^2/\delta^2}{\lambda} T^{2-1/d}}.
\]
Combining with Lemma 2, we can get the bound in Theorem 6.

D.4 Proof of Corollary 6.1

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 \(t, t = 1, \ldots, T\), with a probability of \(1 - \delta, \delta \in (0, 1)\), we have the following lower bound of the cumulative reward
\[
F_T \geq \sum_{t=1}^{T} [MF_t - MR_t - 2h_T T^{3/2} \left( 4d \log \left( 1 + \frac{T}{d\lambda} \right) + \frac{1}{\lambda} \right) - 4 \sqrt{\frac{2}{\lambda}L_{2\phi}^{1/2} T^{1 - \frac{1}{2}}} - 2h_T \sqrt{\frac{2Td}{\lambda}} \log (1 + \frac{T}{d\lambda})],
\]
where \(MR_t\) is the optimal worst-case regret for round \(t\) in Eqn. (11), \(d\) is the dimension of \(x_t\), \(\hat{d}\) is the effective dimension defined in the proof of Proposition 5, \(\hat{d}\) is the rank of \(K_t\), and \(h_T\) is given in Lemma 7.

Proof. By Lemma 1, with a probability \(1 - \delta, \delta \in [0, 1]\), we can bound the reward as below
\[
f(x_t, a^t_t) \geq U_t (x_t, a^t_t) - 2h_t s_t (x_t, a^t_t) + \left( U_t (x_t, A^t_t(x_t)) - U_t (x_t, A^t_t(x_t)) - 2h_t s_t (x_t, a^t_t) \right) = - D_{a^t_t} (x_t) + U_t (x_t, A^t_t(x_t)) - 2h_t s_t (x_t, a^t_t),
\]
Recall that the upper bound of the worst-case degradation is defined as \(D_{a^t_t} = \max_{x \in \mathcal{B}_\Delta(\hat{x}_t)} \left\{ U_t (x_t, A^t_t(x)) - U_t (x_t, a^t_t) \right\}\), so the reward can be further bounded as
\[
f(x_t, a^t_t) \geq \left( - D_{a^t_t} \right) + U_t (x_t, A^t_t(x_t)) - 2h_t s_t (x_t, a^t_t) \geq \left( - D_{a^t_t} \right) + \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x_t, A^t_t(x(t)) - 2h_t s_t (x_t, a^t_t) \geq \left( - D_{a^t_t} \right) + \max_{a \in A} \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x, a) - 2h_t s_t (x_t, a^t_t),
\]
where the last inequality is because \(\min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x, A^t_t(x)) = \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} \max_{a \in A} U_t (x, a) \geq \max_{a \in A} \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x, a)\) by max-min inequality.

Note that \(\max_{a \in A} \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x, a)\) is the arm selection policy of MaxMinUCB whose solutions are \(a^t_t\) and \(x^t_t\). This observation is important since it bridges MinWD and MaxMinUCB. Also, recall that in Eqn. (44), \(a_t = \arg \max_{a \in A} \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} f(x, a)\) is the optimal arm for worst-case reward and \(MF_t = \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} f(x, a_t)\) is the optimal worst-case reward. Then following Eqn. (47), we have
\[
f(x_t, a^t_t) \geq \left( - D_{a^t_t} \right) + U_t (x_t, a^t_t) - 2h_t s_t (x_t, a^t_t) \geq \left( - D_{a^t_t} \right) + U_t (x_t, a^t_t) - 2h_t s_t (x_t, a^t_t) - f(x^t_t, a_t) + MF_t \geq \left( - D_{a^t_t} \right) + MF_t - 2h_t s_t (x_t, a^t_t),
\]
where \(x^t_t = \arg \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} U_t (x, a)\), the second inequality holds by the arm selection strategy of MaxMinUCB such that \(U_t (x_t, a^t_t) \geq U_t (x^t_t, a_t)\), the third inequality comes from the definition of \(MF_t\) in Eqn. (6) which guarantees \(MF_t = \min_{x \in \mathcal{B}_\Delta(\hat{x}_t)} f(x, a_t) \leq f(x^t_t, a_t)\) and the fourth inequality comes from Lemma 2 such that \(U_t (x^t_t, a_t) \geq f(x^t_t, a_t)\).

Finally, since \(\left( - D_{a^t_t} \right) + 2h_t s_t (x_t, a^t_t)\) is the upper bound of instantaneous regret of MinWD, by directly using Theorem 6 we can prove the lower bound reward of MinWD.