Worst-case Performance of Greedy Policies in Bandits with Imperfect Context Observations

Hongju Park and Mohamad Kazem Shirani Faradonbeh

Abstract—Contextual bandits are canonical models for sequential decision-making under uncertainty in environments with time-varying components. In this setting, the expected reward of each bandit arm consists of the inner product of an unknown parameter with the context vector of that arm. The classical bandit settings heavily rely on assuming that the contexts are fully observed, while the study of the richer model of imperfectly observed contextual bandits is immature. This work considers Greedy reinforcement learning policies that take action as if the current estimates of the parameter and of the unobserved contexts coincide with the corresponding true values. We establish that the non-asymptotic worst-case regret grows poly-logarithmically with the time horizon and the failure probability, while it scales linearly with the number of arms. Numerical analysis showcasing the above efficiency of Greedy policies is also provided.

Index Terms—Contextual bandits; Reinforcement learning; Regret analysis; Iterative learning; Statistical learning

I. INTRODUCTION

Contextual bandits are ubiquitous models for sequential decision-making in environments with finite action spaces. The range of applications is extensive and includes different problems for which time-varying and action-dependent information are important, such as personalized recommendation of news articles, healthcare interventions, advertisements, and clinical trials [1], [2], [3], [4], [5], [6], [7].

In many applications, consequential variables for decision-making are not perfectly observed. Technically speaking, in the bandit problem, the context vectors are observed in a partial, transformed, or noisy manner [8], [9], [10]. Furthermore, since perfect observation can be considered as a special case of partially observed contexts, sequential decision-making algorithms for the latter family of problems provide a richer class of settings compared to the former. Accordingly, models with partial observations are commonly used in different problems, including space-state models, robot control, and filtering [11], [12], [13].

We study contextual bandits with partially observed context vectors. The probabilistic structure of the problem under study, as time proceeds, is as follows. At every time step, there are \( N \) available arms, each of which has an unknown context that is denoted by \( x_i(t) \in \mathbb{R}^{d_x} \) for arm \( i \) at time \( t \). The context vectors are generated independently of the previous contexts and the other arms, according to a multivariate normal distribution \( \mathcal{N}(0_{d_x}, \Sigma) \). Moreover, the corresponding observation of \( x_i(t) \) is \( y_i(t) \in \mathbb{R}^{d_y} \), while the stochastic reward \( r_i(t) \) of arm \( i \) is determined by the context vector and the unknown parameter \( \mu_\ast \):

\[
y_i(t) = Ax_i(t) + \zeta(t),
\]

\[
r_i(t) = x_i(t)\mu_\ast + \psi(t).
\]

Above, \( \zeta(t) \) and \( \psi(t) \) are the noises of observation and reward, which are identically distributed and independent following the distributions \( \mathcal{N}(0_{d_y}, \Sigma_y) \) and \( \mathcal{N}(0, \gamma_\ast^2) \), respectively. Further, the known \( d_y \times d_x \) sensing matrix \( A \) captures the relationship between \( x_i(t) \) and the noiseless portion of \( y_i(t) \). The above structure holds for all arms and at all time \( t \). In this model, as compared to the classic contextual bandits, there are uncertainties about the contexts as well as about the parameter \( \mu_\ast \).

At each time, the goal is to learn to choose the optimal arm \( a^\ast(t) \) by utilizing the available information by time \( t \). Note that the optimal arm can change at every time step. An agent chooses an arm based on the previously collected data from the model in (1); \( \{a(\tau)\}_{1 \leq \tau \leq t-1}, \{r_a(\tau)\}_{1 \leq \tau \leq t-1} \) as well as the observations at the time; \( \{y_i(t)\}_{1 \leq i \leq N} \). Then, once the action is taken, the resulting reward of the chosen arm will be provided to the agent according to the equation in (2), while the rewards of the other arms are not observed. Clearly, to choose high-reward arms, the agent needs accurate estimates of the unknown parameter \( \mu_\ast \), as well as those of the contexts \( x_i(t) \), for \( i = 1, \ldots, N \). However, because \( x_i(t) \) is not observed, estimation of \( \mu_\ast \) is only through the observation \( y_i(t) \). Thereby, designing efficient reinforcement learning policies with guaranteed performance is challenging.

Learning strategies for contextual bandits are investigated in the literature, assuming that the context vectors are fully observed. Early papers focus on reinforcement learning policies that utilize Upper-Confident-Bounds (UCB) for addressing the exploitation-exploration trade-off [14], [15], [16]. Another popular and efficient family of policies uses randomized exploration, usually in the (Bayesian) form of Thompson sampling [17], [18], [19], [20]. Recently, for con-
texts generated under certain conditions, it has been shown that exploration-free Greedy policies are efficient [21].

Currently, the study of efficient algorithms with theoretical performance guarantees for imperfect context observations is incomplete. Notably, the imperfectness of observation frequently appears in different areas of applications. The causes are various, including privacy regulations, measurement errors, and missing data [12], [13], [22], [23]. The existing analyses study some special cases, such as those with invertible sensing matrices, where asymptotic results are shown for UCB-type algorithms and Thompson sampling, as well as in presence of additional information [24], [25], [10]. For Greedy algorithms, numerical analyses indicate that they outperform Thompson sampling in partially observed stochastic contextual bandits [26]. Importantly, Greedy policies are of special interest in settings that exploration is not (ethically) permitted, such as precision medicine [21].

We perform the \textit{finite-time worst-case} analysis of Greedy reinforcement learning policies for imperfectly observed contextual bandits. We provide high probability regret bounds that consist of poly-logarithmic factors of the time horizon and of the failure probability. Furthermore, the effects of other quantities such as the number of arms, dimensions, and properties of the noise processes, are fully characterized. Illustrative numerical experiments showcasing the efficiency of Greedy policies are also provided.

Different technical difficulties arise in the analysis of reinforcement learning policies in partially observed contextual bandits. First, one needs to study the eigenvalues of the empirical covariance matrices, since the estimation accuracy heavily depends on them. Furthermore, it is required to consider the number of times the algorithm selects suboptimal arms. Note that both quantities are stochastic and so worst-case (i.e., high probability) results are needed for a statistically dependent sequence of random objects. To obtain the presented theoretical results, we employ advanced technical tools from martingale theory and random matrices. Indeed, by utilizing concentration inequalities for matrices with martingale difference structures, we carefully characterize the effects of order statistics and tail properties of the estimation errors.

The remainder of this paper is organized as follows. In Section II, we formulate the problem and discuss the relevant preliminary materials. Next, a Greedy reinforcement learning policy for contextual bandits with imperfect context observations is presented in Section III. In Section IV, we provide theoretical performance guarantees for the proposed algorithm, followed by numerical experiments in Section V. Finally, we conclude the paper and discuss future directions.

We use $A^\top$ to refer to the transpose of the matrix $A \in \mathbb{C}^{d \times q}$. For a vector $v \in \mathbb{C}^d$, we denote the $\ell_2$ norm by $\|v\| = \left( \sum_{i=1}^{d} |v_i|^2 \right)^{1/2}$. Additionally, $C(A)$ and $C(A)^\perp$ are employed to denote the column-space of the matrix $A$ and its orthogonal subspace, respectively. Further, $P_{C(A)}$ is the projection operator onto $C(A)$, and $\lambda_{\text{min}}(A)$ ($\lambda_{\text{max}}(A)$) denotes the minimum (maximum) eigenvalue of $A$. Besides, $O(\cdot)$ is the order of magnitude such that $f(n) = O(g(n))$ denotes $\limsup_{n \to \infty} |f(n)|/g(n) < \infty$, for a real-valued function $f$ and a strictly positive valued function $g$. Finally, we have $\{X_i\}_{i \in E} = \{X_i : i \in E\}$, and $I(\cdot)$ is the indicator function.

\section{Problem Formulation}

First, we formally discuss the problem of contextual bandits with partially observed contexts. The bandit machine under consideration has $N$ arms, each of which has its own \textit{unobserved} context $x_i(t)$, for $i \in \{1, \ldots, N\}$. Equation (1) presents the observation model, where the observations $\{y_i(t)\}_{1 \leq i \leq N}$ are linearly transformed functions of the contexts, perturbed by additive noise vectors $\{\zeta_i(t)\}_{1 \leq i \leq N}$.

Equation (2) describes the reward generation for different arms, depicting that if the agent selects arm $i$, then the resulting reward is an \textit{unknown} linear function of the unobserved context vector, subject to some additional randomness caused by $\psi_i(t)$.

The agent aims to maximize the cumulative reward over time, by utilizing the sequence of observations. To gain the maximum possible reward, the agent needs to learn the relationship between the reward $r_i(t)$ and the observation $y_i(t)$. For that purpose, we proceed by considering the conditional distribution of the reward $r_i(t)$ given the observation $y_i(t)$.

That is, $P(r_i(t)|y_i(t))$, which is nothing but

$$
N(y_i(t)^T D \mu_s, \gamma_r^2),
$$

where $D = (A^\top \Sigma_x^{-1} A + \Sigma_y^{-1})^{-1} A^\top \Sigma_y^{-1}$ and $\gamma_r^2 = \mu_s^\top (A^\top \Sigma_x^{-1} A + \Sigma_y^{-1})^{-1} \mu_s + \gamma_r^2$.

Based on the conditional distribution in (3), in order to maximize the expected reward given the observation, one can consider the conditional expectation of the reward given the observations; $y_i(t)^T D \mu_s$. So, letting $\eta_s = D \mu_s$ be the transformed parameter, we focus on the estimation of $\eta_s$. The rationale is twofold; first, the conditionally expected reward can be inferred with only knowing $\eta_s$, regardless of the exact value of the true parameter $\mu_s$. Second, $\mu_s$ is not estimable when the rank of the sensing matrix $A$ in the observation model is less than the dimension of $\mu_s$. Indeed, estimability of $\mu_s$ needs the restrictive assumptions of non-singular $A$ and $d_s \leq d_x$.

The optimal policy that reinforcement learning policies need to compete against, knows the true parameter $\mu_s$. That is, to maximize the reward given the output observations, the optimal arm at time $t$, denoted by $a^*(t)$, is

$$
a^*(t) = \arg \max_{1 \leq i \leq N} y_i(t)^\top \eta_s.
$$

Then, the performance degradation due to uncertainty about the environment that the parameter $\mu_s$ represents, is the assessment criteria for reinforcement learning policies. So, we consider the following performance measure, which is commonly used in the literature and is known as \textit{regret} of

\textit{II. Problem Formulation}

First, we formally discuss the problem of contextual bandits with partially observed contexts. The bandit machine under consideration has $N$ arms, each of which has its own \textit{unobserved} context $x_i(t)$, for $i \in \{1, \ldots, N\}$. Equation (1) presents the observation model, where the observations $\{y_i(t)\}_{1 \leq i \leq N}$ are linearly transformed functions of the contexts, perturbed by additive noise vectors $\{\zeta_i(t)\}_{1 \leq i \leq N}$.

Equation (2) describes the reward generation for different arms, depicting that if the agent selects arm $i$, then the resulting reward is an \textit{unknown} linear function of the unobserved context vector, subject to some additional randomness caused by $\psi_i(t)$.

The agent aims to maximize the cumulative reward over time, by utilizing the sequence of observations. To gain the maximum possible reward, the agent needs to learn the relationship between the reward $r_i(t)$ and the observation $y_i(t)$. For that purpose, we proceed by considering the conditional distribution of the reward $r_i(t)$ given the observation $y_i(t)$.

That is, $P(r_i(t)|y_i(t))$, which is nothing but

$$
N(y_i(t)^T D \mu_s, \gamma_r^2),
$$

where $D = (A^\top \Sigma_x^{-1} A + \Sigma_y^{-1})^{-1} A^\top \Sigma_y^{-1}$ and $\gamma_r^2 = \mu_s^\top (A^\top \Sigma_x^{-1} A + \Sigma_y^{-1})^{-1} \mu_s + \gamma_r^2$.

Based on the conditional distribution in (3), in order to maximize the expected reward given the observation, one can consider the conditional expectation of the reward given the observations; $y_i(t)^T D \mu_s$. So, letting $\eta_s = D \mu_s$ be the transformed parameter, we focus on the estimation of $\eta_s$. The rationale is twofold; first, the conditionally expected reward can be inferred with only knowing $\eta_s$, regardless of the exact value of the true parameter $\mu_s$. Second, $\mu_s$ is not estimable when the rank of the sensing matrix $A$ in the observation model is less than the dimension of $\mu_s$. Indeed, estimability of $\mu_s$ needs the restrictive assumptions of non-singular $A$ and $d_s \leq d_x$.

The optimal policy that reinforcement learning policies need to compete against, knows the true parameter $\mu_s$. That is, to maximize the reward given the output observations, the optimal arm at time $t$, denoted by $a^*(t)$, is

$$
a^*(t) = \arg \max_{1 \leq i \leq N} y_i(t)^\top \eta_s.
$$

Then, the performance degradation due to uncertainty about the environment that the parameter $\mu_s$ represents, is the assessment criteria for reinforcement learning policies. So, we consider the following performance measure, which is commonly used in the literature and is known as \textit{regret} of
the reinforcement learning policy that selects the sequence
of actions \( a(t), t = 1, 2, \ldots \):

\[
\text{Regret}(T) = \sum_{t=1}^{T} (y_{a^*(t)}(t) - y_{a(t)}(t))^\top \eta_s. 
\]

In other words, the regret at time \( T \) is the difference in the
sum of rewards obtained up to time \( T \), between the optimal
arms \( a^*(t) \) and the arm \( a(t) \) chosen by the reinforcement
learning policy based on the observations by the time \( t \). Note
that this difference does not directly depend on the unknown
contexts \( \{x_i(t)\}_{1 \leq i \leq N} \) since the optimal policy that selects
\( a^*(t) \) does not observe the context vectors, and decides
merely based on the observations \( y_i(t), \) for \( i = 1, \ldots, N \).

III. REINFORCEMENT LEARNING POLICY

In this section, we explain the details of the Greedy
algorithm for contextual bandits with partially observed
contexts. Although inefficient in some reinforcement learn-
ing problems, Greedy algorithms have a logarithmic regret
bound for contextual bandits with fully observed contexts
under certain conditions including covariate diversity [21].
Intuitively speaking, the latter condition expresses that the
context vectors provide information along all dimensions in
\( \mathbb{R}^{d_x} \) so that additional exploration is not necessary.

As discussed in Section II, it suffices for a policy to learn
the arm that maximizes \( \mathbb{E}[r_i(t)|y_i(t)] = y_i(t)^\top \eta_s \). To that
end, we estimate \( \eta_s \) using the least-squares estimator

\[
\check{\eta}(t) = \arg \min_\eta \sum_{\tau=1}^{t} (r_{a(\tau)}(\tau) - y_{a(\tau)}(\tau)^\top \eta)^2. 
\]

(6)

Then, the Greedy policy works with \( \check{\eta}(t) \) in lieu of the truth
\( \eta_s \). So, the algorithm selects the arm \( a(t) \) at time \( t \), such that
\( a(t) = \arg \max_{1 \leq i \leq N} y_i(t)^\top \check{\eta}(t) \). Thanks to the structure
of the parameter estimates in (6), one can update \( \check{\eta}(t) \) in a
recursive fashion. The recursion relies on updating the empirical
inverse covariance matrix \( B(t) \), according to

\[
B(t + 1) = B(t) + y_{a(t)}(t)y_{a(t)}(t)^\top, \quad \check{\eta}(t + 1) = (B(t)\check{\eta}(t) + y_{a(t)}(t)r_{a(t)}(t)) \quad (7)
\]

In the above recursions, the initial values consist of
\( B(1) = \Sigma^{-1} \), for some arbitrary positive definite square
matrix \( \Sigma \), and \( \check{\eta}(1) = \eta \) for an arbitrary vector \( \eta \) in \( \mathbb{R}^{d_x} \).
Algorithm 1 describes the pseudo-code for the Greedy policy.

Algorithm 1: Greedy policy for contextual bandits with
imperfect context observations

1: Set \( B(1) = \Sigma^{-1} \), \( \check{\eta}(1) = \eta \)
2: for \( t = 1, 2, \ldots \) do
3: Select arm \( a(t) = \arg \max_{1 \leq i \leq N} y_i(t)^\top \check{\eta}(t) \)
4: Gain reward \( r_{a(t)}(t) = x_{a(t)}(t)^\top \mu_s + \psi_{a(t)}(t) \)
5: Update \( B(t + 1) \) and \( \check{\eta}(t + 1) \) by (7) and (8)
6: end for

IV. THEORETICAL PERFORMANCE GUARANTEES

In this section, we present a theoretical analysis for
Algorithm 1 in the previous section. The result provides a
worst-case analysis and establishes a high probability upper-
bound for the regret defined in (5).

Theorem 1. Assume that Algorithm 1 is used in a bandit with
\( N \) arms and the observation dimension \( d_y \). Then, with
probability at least \( 1 - 4\delta \), \( \text{Regret}(T) \) of the order

\[
\text{cond}(A\Sigma_x A^\top + \Sigma_y)\gamma^2 N d_y^{3/2} \left( \log \frac{N d_y T}{\delta} \right)^{5/2} \log \frac{d_y T}{\delta},
\]

where \( \text{cond}(\cdot) \) denotes the condition number of a matrix and \( \gamma^2 \) is the conditional variance of reward defined in (3).

Before proving the theorem, we discuss its implications
and intuitions. The regret bound above scales linearly with
the number of arms \( N \), with \( d_y^{3/2} \) for the dimension of the
observations \( d_y \), and poly-logarithmically with time \( T \).
The dimension of unobserved context vectors does not affect the
regret because the optimal policy in (4) does not have the
exact values of the context vectors. So, in a similar way to
the reinforcement learning policy, the optimal policy needs to
estimate the contexts as well, as \( y_i(t)^\top \eta_s \) in (4) is an estimate of
\( x_i(t)^\top \mu_s \) for the optimal policy to find the optimal arm.

The rationale of the linear growth of the regret bound in
Theorem 1 with the number of arms \( N \) is that for larger \( N \), the policy is more likely to choose one of the
sub-optimal arms, which makes the larger growths of the
regret more likely. In addition, the terms \( d_y^{3/2} \) and \( \text{cond}(A\Sigma_x A^\top + \Sigma_y) \eta_s \) are generated by the \( \ell_2 \) norm
of the stochastic observation vectors, for which the following
high probability upper-bounds hold:

\[
\|y_i(t)\|_2^2 = O(\lambda_{\max}(A\Sigma_x A^\top + \Sigma_y)d_y \log (NTd_y/\delta)).
\]

The poly-logarithmic scaling of the regret bound with \( T \) and \( \delta \) originates from the magnitude of the random vectors
of contexts and observations, as well as from the uncertainty
about the true parameter \( \mu_s \). On the other hand, \( \gamma^2 \) indicates
the role of the reward noise. Simply, if the reward observations
are noisier, it will be harder to learn the optimal arms.

Finally, the ratio of the largest and the smallest eigenvalues
of \( A\Sigma_x A^\top + \Sigma_y \) that constitutes \( \text{cond}(A\Sigma_x A^\top + \Sigma_y) \) in the
regret bound, reflects the effect of the randomnesses of the
observation vectors in different directions.

Proof. We use the following intermediate results, for which
the proofs are delegated to the longer version of this paper
[27]. For simplicity, let \( \tilde{\eta}(t) \) be a random variable with
\( \mathbb{E}[\tilde{\eta}(1)] = \eta_s \) and \( \text{Cov}(\tilde{\eta}(1)) = \Sigma^{-1} \), so that
\( \mathbb{E}[\tilde{\eta}(t)] = \eta_s \) and \( \text{Cov}(\tilde{\eta}(t)|B(t)) = B(t)\gamma^2 \) for all \( t \). In addition, \( S_y \)
denotes \( \text{Cov}(y_i(t)) = A\Sigma_x A^\top + \Sigma_y \) of which smallest and
largest eigenvalues are \( \lambda_{s1} \) and \( \lambda_{s2} \), respectively. To begin
with, for \( 0 < \delta < 0.25 \), we define

\[
W_T = \max_{\{1 \leq \tau \leq t \text{ and } 1 \leq i \leq N\}} \|S_y^{-1/2} y_i(\tau)\|_\infty \leq v_T(\delta), \quad (9)
\]
where $v_T(\delta) = (2\log(Nd_yT/\delta))^{1/2}$. Lemma 1 guarantees that all the observations up to time $T$ are generated in the truncation event $W_T$ with probability at least $1 - \delta$.

**Lemma 1.** For the event $W_T$ defined in (9), we have $P(W_T) \geq 1 - \delta$.

Lemma 2 sets the stage for analyzing the (unnormalized) empirical inverse covariance $B(t)$ in (7).

**Lemma 2.** Let $\xi = \{X_1, \ldots, X_n\}$ be the sigma-field generated by the random objects $X_1, \ldots, X_n$. For the observation $y_{a(t)}(t)$ of the chosen arm at time $t$, the estimator $\hat{\eta}(t)$ defined in (8), and the filtration $\{F_t\}_{1 \leq t \leq T}$ defined according to

$$F_t = \sigma(\{a(\tau)_1 \leq \tau \leq t, \{y(\tau)_1 \leq \tau \leq t, 1 \leq \tau \leq T, \{r_{a(\tau)}(\tau)\}_1 \leq \tau \leq t\}),$$

we have

$$E[V_t | F_{t-1}] = P_{C(S_y^{1/2}\hat{\eta}(t))}(kN - 1) + I_d,$$

where $V_t = S_y^{-1/2}y_{a(t)}(t)\hat{y}_{a(t)}(t)^\top S_y^{-1/2}$ and $kN = E\left[\left(\max_{1 \leq i \leq N}(Z_i)\right)^2\right]$, for $N$ independent $Z_i$ with the standard normal distribution and $S_y = \text{Cov}(y(t))$. That is, $kN$ is the expected maximum of $N$ independent standard normal random variables.

We also use the following concentration inequality for random matrices.

**Lemma 3 (Matrix Azuma Inequality [28]).** Consider the sequence $\{M_k\}_{1 \leq k \leq K}$ of symmetric $d \times d$ random matrices adapted to some filtration $\{F_k\}_{1 \leq k \leq K}$, such that $E[M_k | F_{k-1}] = 0$. Assume that there is a deterministic sequence of symmetric matrices $\{A_k\}_{1 \leq k \leq K}$ that satisfy $M_k^2 \leq A_k^2$, almost surely. Let $\sigma^2 = \|\sum_{1 \leq k \leq K} A_k^2\|$. Then, for all $\varepsilon > 0$, it holds that

$$P\left(\lambda_{\min}\left(\sum_{k=1}^{K} M_k\right) \geq \varepsilon\right) \leq d \cdot e^{-\varepsilon^2/8\sigma^2}.$$

Lemma 4 provides a high probability lower bound for the minimum eigenvalue of $B(t)$. Then, Lemma 5 bounds the estimation error.

**Lemma 4.** For $B(t)$ in (7) and $t \leq T$, on the event $W_T$ defined in (9), with probability at least $1 - \delta$, we have

$$\lambda_{\min}(B(t)) \geq \lambda_{s1}(t - 1) \left(1 - \sqrt{\frac{32v_T(\delta)^4}{t - 1} \log \frac{d_yT}{\delta}}\right).$$

**Lemma 5.** In Algorithm 1, let $\hat{\eta}(t)$ be the parameter estimate as defined in (8). Then, for $t \leq T$, on the event $W_T$ in (9), we have

$$P\left(\|\hat{\eta}(t) - \eta\| > \varepsilon \|B(t)\|\right) \leq 2\exp\left(-\frac{\varepsilon^2}{2d_y\lambda_{\min}(B(t)^{-1})\gamma_y}\right).$$

Next, Lemma 6 gives an upper bound for the probability that Algorithm 1 does not choose the optimal arm at time $t$. Then, Lemma 7 studies the weighted sum of indicator functions $I(a^*(t) \neq a(t))$ that count the effective number of times that the algorithm chooses sub-optimal arms.

**Lemma 6.** Let $\lambda_t = \lambda_{\max}(B(t)^{-1})$. Given $B(t)$, the probability of choosing a sub-optimal arm is bounded as follows:

$$P(a^*(t) \neq a(t)|B(t)) \leq \frac{2Nd_y^{1/2}v_T(\delta))\gamma_y}{\sqrt{\eta^2 S_y \eta}} \lambda_t^{1/2}.$$
with the probability at least $1 - \delta$, for all $t^* < t \leq T$. By Lemma 5 and (13), for all $t^* < t \leq T$, with the probability at least $1 - 3\delta$, we have
\begin{equation}
\lambda_{s_2}^{1/2} v_T(\delta) \|\tilde{h}(t) - \eta_s\| \leq a_1 (t - 1)^{-1/2},
\end{equation}
where $a_1 = 4(\lambda_{s_2}/\lambda_{s_1})^{1/2} \gamma_{T y} v_T(\delta) \sqrt{2 d_y \log(2T \delta^{-1})}$.

Thus, with $(y_{a^*(t)} - y_{a(t)})^\top \eta_s \leq 2\lambda_{s_2}^{1/2} v_T(\delta) \|\eta_s\|$ for $t < t^*$, the regret can be represented
\begin{equation}
\text{Regret}(T) \leq \sum_{t < t^*} 2\lambda_{s_2}^{1/2} v_T(\delta) \|\eta_s\|
+ \sum_{t^* \leq t \leq T} a_1 (t - 1)^{-1/2} I(a^*(t) \neq a(t)),
\end{equation}
with the probability at least $1 - 3\delta$. Now, we consider the probability to choose the optimal arm at time $t$. By Lemma 6, we have
\begin{equation}
\sum_{t^* \leq t \leq T} \mathbb{P}(a^*(t) \neq a(t)|B(t)) \leq 2^{3/2} N \lambda_{s_2}^{1/2} d_y v_T(\delta) \gamma_{T y} \log T / \|\eta_s\| \lambda_{s_1}^{1/2}.
\end{equation}

(16)

Now, we construct an upper bound for the indicator function $I(a^*(t) \neq a(t))$ in (12), by Lemma 7.
\begin{equation}
\sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t - 1}} I(a^*(t) \neq a(t)) \leq \sqrt{32 \log T \log(T \delta^{-1})}
+ \sum_{t^* \leq t \leq T} \frac{1}{\sqrt{t - 1}} \mathbb{P}(a^*(\tau) \neq a(\tau)|B(\tau)),
\end{equation}
with the probability at least $1 - \delta$. Therefore, by (16) and (17), with the probability at least $1 - 4\delta$, the following inequalities hold for the regret, which yield to the desired result:
\begin{equation}
\text{Regret}(T) = \sum_{t = 1}^{T} (y_{a^*(t)} - y_{a(t)})^\top \eta_s I(a^*(t) \neq a(t))
\leq 2\lambda_{s_2}^{1/2} v_T(\delta) \|\eta_s\| + \sum_{t^* \leq t \leq T} a_1 \frac{1}{\sqrt{t - 1}} I(a^*(t) \neq a(t))
= O\left(\frac{\lambda_{s_2}}{\lambda_{s_1}} \gamma_{T y}^2 N d_y^2/3 \left(\log \frac{N d_y T}{\delta}\right)^{5/2} \log \frac{d_y T}{\delta}\right).
\end{equation}

Finally, using $\text{cond}(S_y) = \lambda_{s_2}/\lambda_{s_1}$, with probability at least $1 - 4\delta$ we have
\begin{equation}
\text{Regret}(T) = O\left(\text{cond}(S_y)^{1/2} \gamma_{T y}^2 N d_y^{3/2} \left(\log \frac{N d_y T}{\delta}\right)^{5/2} \log \frac{d_y T}{\delta}\right).
\end{equation}

(19)

V. NUMERICAL ILLUSTRATIONS

In this section, we perform numerical analyses for the theoretical result in the previous section. We simulate cases for $N = 10, 20, 50$ arms and different dimensions of the observations $d_y = 5, 20, 50$ with a fixed context dimension $d_x = 20$. Each case is repeated 100 times and the average and worst quantities amongst all 100 scenarios are reported.

In Figure 1, the left plot depicts the average-case (solid) and worst-case (dashed) regret among all scenarios, normalized by $\log t$. The number of arms $N$ varies as shown in the graph, while the dimension is fixed to $d_y = 10$. Next, the plot on the right illustrates that the normalized regret increases for different $d_y$ and for the fixed number of arms $N = 5$.

In both plots, the worst-case regret curves are well above the average ones, while the slopes of curves for both cases become flat as time goes on, implying that the worst-case regret grows logarithmically in terms of $t$ as well.

Figure 2 presents the worst-case (normalized) regret at time $T = 2000$. We evaluate the regret by changing the dimension of the context vectors and the observations. To reflect the effect of the dimension, the regret is normalized by the other factors. The plot represents that regret grows linearly with the dimension of the observations.

VI. CONCLUSION

This work investigates reinforcement learning algorithms for contextual bandits where the contexts are partially observed, focusing on the theoretical analysis and provable regret bounds. We establish a high probability regret bound for Greedy algorithms, which grows poly-logarithmically with the time horizon.

There are multiple interesting future directions introduced in this paper. First, it will be of interest to study reinforcement learning policies for settings where each arm has its own parameter. Further, regret analysis for contextual bandits under imperfect context observations where the other parameters such as the covariance matrices of contexts and observations and the sensing matrix are unknown, is another problem for future work.

REFERENCES

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

[2] D. Bouneffouf, A. Bouzeghoub, and A. L. Gançarski, “A contextual-bandit algorithm for mobile context-aware recommender system,” in International conference on neural information processing . Springer, 2012, pp. 324–331.

[3] A. Tewari and S. A. Murphy, “From ads to interventions: Contextual bandits in mobile health,” in Mobile Health . Springer , 2017, pp. 495–517.

[4] I. Nahum-Shani, S. N. Smith, B. J. Spring, L. M. Collins, K. Witkiewitz, A. Tewari, and S. A. Murphy, “Just-in-time adaptive interventions (jitas) in mobile health: key components and design principles for ongoing health behavior support,” Annals of Behavioral Medicine , vol. 52, no. 6, pp. 446–462, 2018.

[5] A. Durand, C. Achilleos, D. Iacovides, K. Strati, G. D. Mitsis, and J. Pineau, “Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis,” in Machine learning for healthcare conference . PMLR, 2018, pp. 67–82.

[6] Y. Varatharajah, B. Berry, S. Koyejo, and R. Iyer, “A contextual-bandit-based approach for informed decision-making in clinical trials,” arXiv preprint arXiv:1809.00258, 2018.

[7] Z. Ren and Z. Zhou, “Dynamic batch learning in high-dimensional sparse linear contextual bandits,” arXiv preprint arXiv:2008.11918, 2020.
Fig. 1: Plots of $\text{Regret}(t)/\log t$ over time for the different number of arms $N = 10, 20, 50$ and $d_y = 5, 20, 50$. The solid and dashed lines represent the average-case and worst-case regret curves, respectively.

Fig. 2: Plot of worst-case normalized regret at $T = 2000$ versus the dimension of the contexts ($d_x = 10, 20, 40, 80$), for different dimensions of the observations $d_y = 10, 20, 40, 80$.

[8] A. Bensoussan, *Stochastic control of partially observable systems*. Cambridge University Press, 2004.

[9] D. Bourneffouf, I. Rish, G. A. Cecchi, and R. Féraud, “Context attentive bandits: Contextual bandit with restricted context,” *arXiv preprint arXiv:1705.03821*, 2017.

[10] T. Tennenholtz, U. Shalit, S. Mannor, and Y. Efroni, “Bandits with partially observable confounded data,” in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2021.

[11] I. Nagrath, *Control systems engineering*. New Age International, 2006.

[12] J.-W. Lin, C.-W. Chen, and C.-Y. Peng, “Kalman filter decision systems for debris flow hazard assessment,” *Natural hazards*, vol. 60, no. 3, pp. 1255–1266, 2012.

[13] Y. Kang, C. Roh, S.-B. Suh, and B. Song, “A lidar-based decision-making method for road boundary detection using multiple kalman filters,” *IEEE Transactions on Industrial Electronics*, vol. 59, no. 11, pp. 4360–4368, 2012.

[14] P. Auer, “Using confidence bounds for exploitation-exploration trade-offs,” *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[15] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” *Advances in neural information processing systems*, vol. 24, pp. 2312–2320, 2011.

[16] W. Chu, L. Li, L. Reyzin, and R. Schapire, “Contextual bandits with linear payoff functions,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 208–214.

[17] O. Chapelle and L. Li, “An empirical evaluation of thompson sampling,” *Advances in neural information processing systems*, vol. 24, pp. 2249–2257, 2011.

[18] S. Agrawal and N. Goyal, “Thompson sampling for contextual bandits with linear payoffs,” in *International Conference on Machine Learning*. PMLR, 2013, pp. 127–135.

[19] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, “On adaptive linear–quadratic regulators,” *Automatica*, vol. 117, p. 10892, 2020.

[20] A. Modi and A. Tewari, “No-regret exploration in contextual reinforcement learning,” in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 829–838.

[21] H. Bastani, M. Bayati, and K. Khosravi, “Mostly exploration-free algorithms for contextual bandits,” *Management Science*, vol. 67, no. 3, pp. 1329–1349, 2021.

[22] H. Sheth and R. Younes, “Review of optimization methods for cancer chemotherapy treatment planning,” *Journal of Computer Science & Systems Biology*, vol. 8, no. 2, p. 74, 2015.

[23] I. Azimi, T. Pahikkala, A. M. Rahmani, H. Niela-Vilén, A. Axelin, and P. Liljeberg, “Missing data resilient decision-making for healthcare iot through personalization: A case study on maternal health,” *Future Generation Computer Systems*, vol. 96, pp. 297–308, 2019.

[24] S.-Y. Yun, J. H. Nam, S. Mo, and J. Shin, “Contextual multi-armed bandits under feature uncertainty,” *arXiv preprint arXiv:1703.01347*, 2017.

[25] H. Park and M. K. S. Faradonbeh, “Analysis of thompson sampling for partially observable contextual multi-armed bandits,” *IEEE Control Systems Letters*, vol. 6, pp. 2150–2155, 2021.

[26] ——, “Efficient algorithms for learning to control bandits with unobserved contexts,” *arXiv preprint arXiv:2202.00867*, 2022.

[27] ——, “Worst-case performance of greedy policies in bandits with imperfect context observations,” *arXiv preprint arXiv:2204.04773*, 2022.

[28] J. A. Tropp, “User-friendly tail bounds for sums of random matrices,” *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.