Online Learning with Regularized Knowledge Gradients

Donghun Lee[1]* and Warren B. Powell[2]

[1] Korea University, Seoul 02841, South Korea
holy@korea.ac.kr
[2] Princeton University, Princeton NJ 08544, USA
wbpowell328@gmail.com

Abstract. We introduce a simple and effective regularization of knowledge gradient (KG) and use it to present the first sublinear regret bound result for KG-based algorithms. We construct online learning with regularized knowledge gradients (ORKG) algorithm with independent Gaussian belief model, and prove that ORKG algorithm achieves sublinear regret upper bound with high probability facing bounded independent Gaussian multi-armed bandit (MAB) problems. The theoretical properties of regularized KG and ORKG algorithm are analyzed, and the empirical characteristics of ORKG algorithm are empirically validated with MAB benchmark simulations. ORKG algorithm shows top-tier performance comparable to select MAB algorithms with provable regret bounds.

Keywords: Knowledge Gradient · Online Learning · Regret Analysis

1 Introduction

This paper considers the problem of making best possible decisions facing uncertainty, in which a decision-making agent repeatedly chooses from a set of decisions and then observes an outcome from which a bounded quantifiable reward can be derived. We assume that the agent knows the set of possible decisions, which is finite and remains the same over the time horizon in which the agent choose and learns. If such an agent is evaluated on how well it finds out which choice incurs the best reward, disregarding the rewards incurred by its choices while learning, the agent is facing a ranking and selection (R&S) problem.

Knowledge gradient (KG) is an algorithm proposed to solve R&S problems with independent Gaussian model assumption [6], and later with different assumptions such as correlated Gaussian model [7], Gaussian process model [17], binary cost function [24], locally nonlinear parametric models [11], and repeated noisy measurements [10]. Empirical effectiveness of KG-based algorithms has

* This work is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1G1A1102828) and the Korea University grant (No. K2120891)

been demonstrated in diverse fields where R&S problems can be applied: for example, drug discovery [14], chemical engineering [5], fleet management [23, 12], COVID responses [21], and clinical trials [22].

However, R&S problem disregards the reward incurred by the choices made by the agent while it is learning. As such, R&S problem is ill-suited to model online learning problems, in which every single reward incurred by the agent counts, and 2) the remaining number of choices the agent must make may be unknown. Little work has been done to utilize KG in online learning, where a most notable approach assuming the agent knows the remaining number of choices [15, 16]. In this paper, we present a new approach to utilize KG to solve online learning problem *with unknown time horizon*.

Novel contribution of this manuscript is summarized as follow. We present Online learning with Regularized Knowledge Gradients (ORKG) algorithm with independent Gaussian belief, a novel online learning algorithm that uses knowledge gradient. We provide theoretical analysis of ORKG, including the proof of ORKG's regret upper bound of $O(\sqrt{KT\ln(KT)})$ in stochastic MAB problems with K bounded independent Gaussian arms, which is the first sublinear regret bound for knowledge gradient based algorithms. We also perform empirical validation of the theoretical properties of ORKG and empirical sensitivity analysis of the key hyperparameters of ORKG. Lastly, we verify empirical performance of ORKG in Gaussian stochastic MAB problems against other well-known MAB algorithms with provable regret bounds.

2 Problem Setting

We consider "online" sequential decision problem in which a decision-making agent faces partial information stochastic MAB problem, in particular with bounded Gaussian stochastic arms and unknown total number of decisions to make. For each time index $t \in \{0, 1, \cdots, T-1\}$ with unknown finite time horizon T, the agent must make a decision, denoted by x, among $K < \infty$ mutually independent arms that can be indexed by $i \in \{1, 2, \cdots, K\}$, and then observe a bounded random reward/contribution C_t from respective arm's distribution with mean μ^i and standard deviation σ^i that are unknown to the agent. We use x_t for decision made at time t, and $\mathcal{X}$ as the set containing all possible decisions. Hence, $\forall t : x_t \in \mathcal{X}$, and $|\mathcal{X}| = K$.

The goal of the agent is twofold: 1) to learn the best arm i^* whose reward distribution has the largest mean (i.e. $\mu^{i^*} = \max_i \{\mu^i\} =: \mu^*$ using the observations incurred by past decisions, and 2) to control the impact of inevitable suboptimality caused by choosing arms that are not the best arm without knowing the best arm *a priori*. Note that the term "online" is *not* the same as in online convex optimization, but instead is related to the second aspect of the goal of the learning agent – that the performance of the agent while it is learning (i.e. "online") is important, as opposed to batch learning such as R&S problems where only final performance matters.

The belief state B_t at time t, for KG-based algorithms with independent Gaussian belief model, is defined as the sufficient information to model Gaussian rewards incurred by each action $x \in \mathcal{X}$. Hence, we define $B_t := \{(\bar{\mu}_t^x, \bar{\sigma}_t^x) | x \in \mathcal{X}\}$, as the set of mean parameter estimates $\bar{\mu}_t^x$ and standard deviation parameter estimates $\bar{\sigma}_t^x$ for all $x \in \mathcal{X}$.

Under independent Gaussian belief model, KG of choosing x at time t can be efficiently computed [8] using the following closed form formula:

$$\nu_t^{KG,x} := \tilde{\sigma}_t^x \left(\xi_t^x \Phi\left(\xi_t^x\right) + \phi\left(\xi_t^x\right) \right), \tag{1}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and the probability density function of standard Gaussian distribution, respectively. ξ_t^x is defined as:

$$\xi_t^x := -\frac{\left| \bar{\mu}_t^x - \max_{x' \neq x} \bar{\mu}_t^{x'} \right|}{\tilde{\sigma}_t^x}, \tag{2}$$

where $\tilde{\sigma}_t^x := \bar{\sigma}_t^x / \sqrt{1 + \left(\sigma^\epsilon / \bar{\sigma}_t^x\right)^2}$. σ^ϵ is the standard deviation of the zero-mean Gaussian measurement noise assumed to be found on all observed reward $C(x)$ for all $x \in \mathcal{X}$. Most KG-based algorithms have σ^ϵ as a hyperparameter.

Using KG as-is to solve online learning problems is expected to fail, because R&S problem disregards the rewards caused by a fixed, known number of choices which it considers as the learning process. From this perspective, KG algorithm for online learning problems (OKG) is proposed [15]. OKG algorithm chooses action x_t at time t as:

$$x_t = \begin{cases} \arg\max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x + (T-t)\, \nu_t^{KG,x} \right\} & (t < T) \\ \arg\max_{x \in \mathcal{X}} \left\{ \bar{\mu}_t^x \right\} & (t \geq T) \end{cases}, \tag{3}$$

where T is the total number of choices to make in the online learning problem. Naturally, OKG algorithm requires knowing the true time horizon T, after which it exploits learned information and choose the action with best expected mean reward.

3 Online Learning with Regularized KG

We present Online learning with Regularized KG (ORKG) with independent Gaussian belief, a novel online learning algorithm with knowledge gradient, in Algorithm 1. Compared to OKG algorithm [15], ORKG introduces two key innovations: 1) standardizing and regularizing knowledge gradient; 2) adaptively learning exploration parameter ρ_t. ORKG contains two key hyperparameters $\kappa_R > 0$ and $0 < \delta < 1$, and we use $\kappa_R = 0.01$, $\delta = 0.01$ as their default values. These hyperparameters are explained in theoretical analysis of ORKG (section 4) and their default values are justified in empirical sensitivity analysis of ORKG (section 5.2).

Algorithm 1 ORKG with Independent Gaussian Belief

1: Initialize belief state: $\{\bar{\mu}_0^x, \bar{\sigma}_0^x\}_{x\in\mathcal{X}}$
2: **for** $t = 0, 1, 2, \cdots$ **do**
3: Compute standardized KG: $\kappa_t^x \leftarrow \frac{\nu_t^{KG,x}}{\bar{\sigma}_t^x}$ ▷ Compute $\nu_t^{KG,x}$ by (1)
4: Compute regularized KG: $\nu_t^{RKG,x} \leftarrow \bar{\sigma}_t^x \max(\kappa_R, \kappa_t^x)$
5: Compute coefficient $\rho_t \leftarrow \sqrt{2\ln\left(\frac{2|\mathcal{X}|}{\delta\pi_t}\right)}\frac{1}{\max\left\{\kappa_R, \min_{x\in\mathcal{X}}\kappa_t^x\right\}}$
6: Choose action: $x_t \leftarrow \arg\max_{x\in\mathcal{X}}\left\{\bar{\mu}_t^x + \rho_t\nu_t^{RKG,x}\right\}$
7: Observe $C_{t+1} \sim C(x_t)$
8: Update $\bar{\mu}_{t+1}^x, \bar{\sigma}_{t+1}^x$ for $x = x_t$ using observation C_{t+1} ▷ Use update rules in [8]

As in step 6 of Algorithm 1, ORKG algorithm chooses action at time t as:

$$x_t = \arg\max_{x\in\mathcal{X}}\left\{\bar{\mu}_t^x + \rho_t\nu_t^{RKG,x}\right\}, \tag{4}$$

where $\rho_t := \sqrt{2\ln\left(\frac{2|\mathcal{X}|}{\delta\pi_t}\right)}\frac{1}{\max\{\kappa_R, \min_{x\in\mathcal{X}}\kappa_t^x\}}$, in which $\delta \in (0,1)$ and π_t is a sequence satisfying $\sum_t^\infty \pi_t = 1$, for example, $\pi_t := \frac{1}{(t+1)^2}\frac{6}{\pi^2}$. With this ρ_t, ORKG balances the exploitation action to maximize $\bar{\mu}_t^x$, the current estimate of mean reward incurred by action x and the exploration action to maximize $\nu_t^{RKG,x}$, the regularized knowledge gradient of action x at time t. It is notable that ORKG does not need to know the time horizon T; whereas OKG algorithm explicitly requires knowing the true T as shown in its decision rule (3). This property allows ORKG to be easily applied to online learning problems where explicit end-of-horizon is unknown or changes over time.

With carefully constructed decision rule, ORKG controls the exploration-exploitation dilemma in online learning problem with unknown horizon, and achieves sublinear regret upper bound as shown in Theorem 1.

Theorem 1. *In stochastic MAB problems with bounded independent Gaussian arms, ORKG algorithm with independent Gaussian belief has regret upper bound:*

$$R_T \leq_p \sqrt{8|\mathcal{X}|T\ln\left(\frac{2|\mathcal{X}|T}{\delta\pi_{T-1}}\right)}L^{RKG}\sigma^\epsilon,$$

with probability $1-\delta$*, where* $0 < \delta < 1$*, and* $L^{RKG} < \infty$ *is a constant uniformly bounding smoothness of regularized KG surface.*

Our proof strategy, which is inspired from GP-UCB algorithm [19], is as follow: first, the deviations of Gaussian rewards are taken with union bounds to bound squared one-step regret with high probability, given δ and κ_R, and then we sum up one-step regrets and bound the regret $R(T)$ and derive ρ_t shown in Algorithm 1. The smoothness constant L^{RKG} is analyzed in greater detail in section 4.2, and complete proof of Theorem 1 is given in appendix A.7.

Therefore, ORKG algorithm with independent Gaussian belief has a sublinear regret upper bound of $O\left(\sqrt{|\mathcal{X}|\,T\ln|\mathcal{X}|\,T}\right)$ with probability $1-\delta$, when its modeling assumption matches the problem specification.

4 Theoretical Analysis

4.1 Regularization of Knowledge Gradient in ORKG

In this section, we define the regularization of KG used in ORKG algorithm, and analyze the theoretical property of the regularized KG on which the sublinear regret bound of ORKG depends.

Conceptual summary of the regularization of KG in ORKG algorithm is as follows: 1) "standardize" KG into a unitless value, 2) force it to have a fixed uniform bound from below, 3) then give back its unit to match KG. Step 1 is achieved by computing standardized KG, and steps 2 and 3 are done in computing regularized KG from standardized KG.

Definition 1. *κ_t^x, standardized knowledge gradient of an action $x \in \mathcal{X}$ at time t is defined for all $x \in \mathcal{X}$ as:*

$$\kappa_t^x := \frac{\nu_t^{KG,x}}{\bar{\sigma}_t^x}, \tag{5}$$

where knowledge gradient $\nu_t^{KG,x}$ is computed from belief state B_t.

κ_t^x is "standardized" KG, in a sense that it has the same unit as ξ_t^x:

$$\kappa_t^x = \underbrace{\frac{\bar{\sigma}_t^x}{\sqrt{\left(\bar{\sigma}_t^x\right)^2 + \left(\sigma^\epsilon\right)^2}}}_{\text{unitless}} \underbrace{\left(\xi_t^x \Phi\left(\xi_t^x\right) + \phi\left(\xi_t^x\right)\right)}_{\text{same unit as } \xi_t^x}, \tag{6}$$

where ξ_t^x is as defined in (2), Φ is the cumulative distribution function, and ϕ is the probability density function of standard normal distribution.

We introduce the following regularization method, designed to achieve a needed property for a sublinear upper bound of the regret of ORKG, and at the same time easy to interpret.

Definition 2. *$\nu_t^{RKG,x}$, the regularized KG for making a decision x at time t given belief state B_t, is defined as*

$$\nu_t^{RKG,x} := \bar{\sigma}_t^x \max\left\{\kappa_R, \kappa_t^x\right\}, \tag{7}$$

where $\kappa_R > 0$ is the regularizing parameter, which is a small arbitrary constant uniform lower bound on κ_t^x for all x, t, and κ_t^x is standardized KG computed at time t given belief state B_t according to Definition 1.

Note that from this regularization originates κ_R, one of the two hyperparameters of ORKG algorithm. κ_R stands for the uniform lower bound on how small κ_t^x can get for all x, t.

4.2 Smoothness of Regularized KG Surface

In ORKG algorithm facing stochastic MAB with finite number of bounded Gaussian independent arms, $\nu_t^{KG,x}$ can be efficiently computed for all $x \in \mathcal{X}$ given B_t. To represent the "surface" of KG with respect to x at t, we consider $\nu_t^{KG} = \left[\nu_t^{KG,1}, \nu_t^{KG,2}, \cdots, \nu_t^{KG,K}\right]$ as a piecewise linear function measured at $x = 1, 2, \cdots, K$. We define a smoothness constant for the surface of KG as:

Definition 3. *$L_t^{KG,x}$, the smoothness constant of KG for action x at time t, is defined as:*

$$L_t^{KG,x} := \frac{\nu_t^{KG,x}}{\min_{x' \in \mathcal{X}} \nu_t^{KG,x'}} \,. \tag{8}$$

$L_t^{KG,x}$ represents the worst case relative difference between KG of x at t and smallest KG across all x at t, up to permutation of $\mathcal{X}$, in the unit of the value of smallest KG at t. It has trivial lower bound of 1, and upper bound of ∞ at $t \to \infty$, suggesting that the KG "surface" may have a very sharp point.

On the other hand, the surface of regularized KG, whose smoothness constant is shown in Definition 4, has a smoothness bound as shown in Lemma 1.

Definition 4. *$L_t^{RKG,x}$, the smoothness constant of regularized KG for action x at time t, is defined, analogous to that of KG (Definition 3), as:*

$$L_t^{RKG,x} := \frac{\max\left\{\kappa_R, \kappa_t^x\right\}}{\max\left\{\kappa_R, \min_{x' \in \mathcal{X}} \kappa_t^{x'}\right\}} \,. \tag{9}$$

Lemma 1. *There exists a finite constant L^{RKG} such that*

$$L_t^{RKG,x} \leq L^{RKG} < \infty \quad \forall x \in \mathcal{X}, \forall t \in \{0, 1, \cdots\} \,. \tag{10}$$

Existence of a constant L^{RKG} is needed to establish the sublinear regret upper bound of ORKG, as the constant appears in the regret bound in Theorem 1. We provide the proof of Lemma 1 in appendix A.3.

5 Empirical Verification

In this section, we present multifaceted empirical verification of the performance of ORKG algorithm in online learning. We use Python package `smpybandit` [3] to implement all stochastic multi-armed bandit (MAB) benchmarks, on an AMD Ryzen 3900x CPU with 64GB of RAM. Benchmarks are randomized and repeated 100 times, and the sample mean and standard deviation from all repeats are reported. For each benchmark scenario, the best result in sample mean and all runner-up results within 1 standard deviation of the best result are boldfaced.

5.1 ORKG Compared to Other KG Algorithms

First, we demonstrate how the theoretical improvements of ORKG is realized, by comparing empirical performance of KG based algorithms in bounded Gaussian stochastic MAB problems. We compare ORKG algorithm against KG with independent Gaussian belief algorithm (KG) [8] and KG for general class of online learning problems algorithm (OKG) [15]. We also test ϵ-greedy algorithm with constant $\epsilon(t) = 0.01$ as a widely known benchmark algorithm frequently seen in applications. The key differences of the algorithms are outlined in Table 1. We use $\sigma_\epsilon = 0.1$ as the value of the common hyperparameter among the KG algorithms for fair comparison.

Table 1: Comparison of algorithms used in section 5.1

	Decision Rule	Belief State	Hyperparameters	Regret Bound
ϵ-greedy	$\bar{\mu}_t^x$ w.p. $1-\epsilon$	$\bar{\mu}_t^x$	$\epsilon(t)$	N/A
KG	$\nu_t^{KG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	σ_ϵ	N/A
OKG	$\bar{\mu}_t^x + (T-t)\nu_t^{KG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	σ_ϵ, T	N/A
ORKG	$\bar{\mu}_t^x + \rho_t \nu_t^{RKG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	$\sigma_\epsilon, \delta, \kappa_R$	$O\left(\sqrt{\lvert\mathcal{X}\rvert T \ln \lvert\mathcal{X}\rvert T}\right)$

We test the algorithms on the stochastic MAB benchmark problems with 5, 10, and 20 arms generating Gaussian rewards, whose mean parameter μ_x sampled equally distanced in $[-5, 5]$, with low variance scenario of $\sigma_x^2 = 0.1$ and high variance scenario $\sigma_x^2 = 1$ for all actions x. For each algorithm, we sum up observed regrets from $t = 1, \cdots, 10000$, and report their mean and standard deviations from 100 independent repeats in Table 2. ORKG shows expected behavior of controlling the cumulative regret throughout all tested settings, whereas other KG algorithms without sublinear regret bounds mostly show large regret. OKG, even when provided with additional information on the

Table 2: Cumulative Regrets in Gaussian Stochastic MAB. Lower is Better.

MAB Setting		Algorithms			
Arms	Variance	ORKG	OKG	KG	ϵ-greedy
5	High	$\mathbf{215 \pm 102}$	$\mathbf{204 \pm 96}$	33100 ± 256	8830 ± 8200
5	Low	$\mathbf{17 \pm 9}$	$\mathbf{12 \pm 12}$	33200 ± 235	6570 ± 8110
10	High	$\mathbf{1060 \pm 85}$	2580 ± 3210	39600 ± 355	14700 ± 11100
10	Low	$\mathbf{40 \pm 9}$	1020 ± 2840	40600 ± 241	17400 ± 11900
20	High	$\mathbf{2210 \pm 105}$	5950 ± 3900	39900 ± 774	19600 ± 10500
20	Low	$\mathbf{96 \pm 10}$	6210 ± 4690	44400 ± 264	21100 ± 14500

true time horizon $T = 10000$, achieves results comparable to ORKG only in 5 arms setting, not in the harder settings with 10 and 20 arms. KG, intended to solve R&S problem, shows worst performance in terms or regrets as expected. Note that ϵ-greedy, a widely used algorithm in practice, shows extremely large standard deviation, suggesting hit-or-miss performance in online learning.

5.2 Sensitivity Analysis of ORKG

ORKG introduces new hyperparameters δ and κ_R compared to other KG algorithms as shown in Table 1. Since those hyperparameters play critical role in the sublinear regret bound of ORKG, we analyze empirical sensitivity of ORKG to δ and κ_R, one by one, tested on Gaussian MAB benchmarks. In the main paper, we present results with 10 arms and high variance only, and full results are found in appendix (Figures B.4 and B.5).

First, we vary $\kappa_R \in \{0.0001, 0.001, 0.01, 0.1, 1\}$ while fixing $\delta = 0.01$, and report the time evolution of cumulative regret against t, averaged over 100 repeats, as trajectories shown in Figure 1.

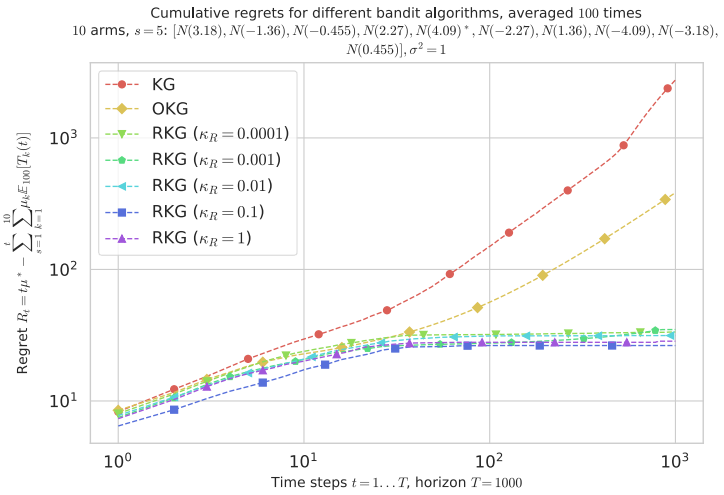

Fig. 1: Sensitivity of ORKG to κ_R in Gaussian MAB ($K = 10, \sigma^2 = 1$, $\delta = 0.01$).

It is evident that ORKG shows robust regret controls regardless of wide range of κ_R, and retains robust advantage over KG and OKG. Considering the intuitive role of κ_R in ORKG to enforce the lower bound of KG and regularizes the smoothness of the KG surface, the subtle sensitivity to κ_R is theoretically expected, and can be interpreted as follows: changing κ_R can change the values of the ORKG decision rule (4) transiently when exploration happens, visualized as minor difference in early-stage trajectories ($T < 10^3$) of ORKGs with different κ_R values in Figure 1.

We recommend the default value of $\kappa_R = 0.01$, based on theoretical understanding of the value should be small enough to become the lower bound of KG, as the intuitive meaning of KG is the expected improvement from a single reward. Also, κ_R can be tuned with *a priori* information or at problem formulation stage: if the gap between the largest mean and the smallest mean of the rewards are known or can be enforced by clipping rewards, then κ_R can be set to be at least sufficiently smaller than the gap.

Next, we vary $\delta \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ while fixing $\kappa_R = 0.01$, and report the time evolution of cumulative regret against t, averaged over 100 repeats, as the trajectories shown in Figure 2.

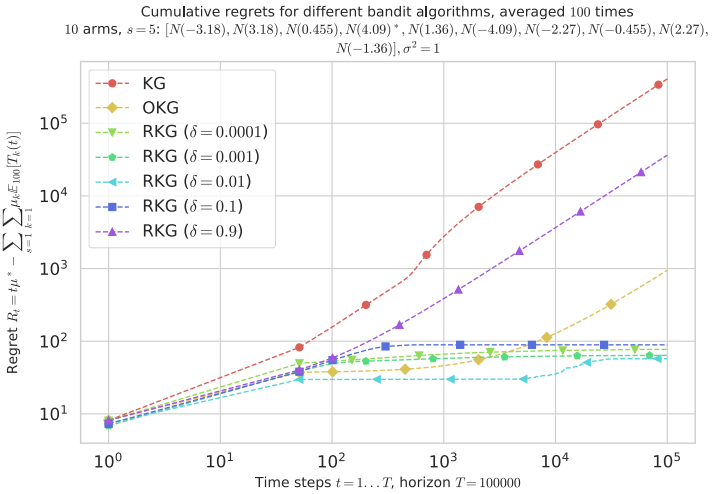

Fig. 2: Sensitivity of ORKG to δ in Gaussian MAB ($K = 10, \sigma^2 = 1$, $\kappa_R = 0.01$).

It is notable that δ affects the behavior of ORKG in mid-range $10^2 < T < 10^4$ to vary, and in most cases, the impact appears to be transient as the regret is controlled for $\delta \leq 0.1$ cases. Drastically different behavior of ORKG is observed for $\delta = 0.9$ case, and this is expected according to the role of δ in ORKG: the probability δ of encountering a reward deviates more than the estimated mean plus exploration bonus term scaled by ρ_t (as given in (4)). Intuitively, larger δ makes ORKG more cautious before greedily exploiting, since δ is the probability of a rare event of facing unexpected rewards after choosing the action according to ORKG decision rule (4), and this is empirically shown by ORKG with $\delta = 0.9$ case in Figure 2. Therefore, it is reasonable to set δ in ORKG as a relatively small value even if $\delta \in (0, 1)$ is theoretically allowed, as δ adjusts how much ORKG should expect the rare events would happen. We recommend the default value of $\delta = 0.01$, as 1% appears to be a good reference point for encountering "rare" events; if more frequent surprises are expected, larger δ is recommended.

5.3 ORKG Performance Validation Against Other MAB Algorithms

We validate empirical performance of ORKG against other MAB algorithms with provable regret bounds, on stochastic Gaussian MAB benchmark problems set up the same way as described in section 5.1. Both classic algorithms and cutting-edge algorithms for MAB are compared against ORKG in this validation, with abbreviated names as follow: UCB [13], kl-UCB [9], EXP3++ [18], TS [20], and BG [4]. Detailed rationale of choosing these algorithms are given in appendix section B.1. For each algorithm, we sum up observed regrets from $t = 1, \cdots, 10000$, and report their mean and standard deviations from 100 independent repeats in Table 3.

Table 3: Cumulative Regrets in Gaussian Stochastic MAB. Lower is Better.

MAB Setting		Algorithms					
Arms	Variance	ORKG	UCB	kl-UCB	TS (G)	EXP3++	BG
5	High	$\mathbf{215 \pm 102}$	$\mathbf{247 \pm 90}$	573 ± 2320	$\mathbf{246 \pm 94}$	1090 ± 113	$\mathbf{235 \pm 105}$
5	Low	$\mathbf{17 \pm 9}$	$\mathbf{30 \pm 10}$	$\mathbf{15 \pm 10}$	$\mathbf{41 \pm 37}$	919 ± 67	$\mathbf{36 \pm 12}$
10	High	$\mathbf{1060 \pm 85}$	$\mathbf{1060 \pm 88}$	1920 ± 2590	1420 ± 698	2920 ± 198	$\mathbf{1070 \pm 99}$
10	Low	$\mathbf{40 \pm 9}$	75 ± 11	$\mathbf{41 \pm 10}$	644 ± 1240	1920 ± 138	85 ± 12
20	High	$\mathbf{2210 \pm 105}$	$\mathbf{2260 \pm 68}$	3240 ± 1930	4590 ± 2210	5480 ± 212	$\mathbf{2240 \pm 72}$
20	Low	$\mathbf{96 \pm 10}$	182 ± 9	$\mathbf{91 \pm 10}$	3010 ± 2490	3470 ± 226	181 ± 12

As shown by boldfaced results across all scenarios, ORKG reliably performs well in all tested Gaussian MAB benchmark scenarios, with the cumulative regret of ORKG is on par with the top-performing algorithm within each scenario; whereas other algorithms show some scenario preferences in which they perform well. Both UCB, a classic algorithm, and Boltzmann-Gumbel (BG), a cutting edge algorithm are the runner-ups, closely followed by kl-UCB, an improved UCB with tighter bound that shows scenario preference different from UCB. We conjecture that the unexpectedly poor performance of EXP3++ may be an unwanted artifact of general-purposing EXP3 algorithm that is originally designed for adversarial MAB problems to have sublinear regrets for stochastic MAB problems as well. Thompson sampling (TS) also show unexpectedly poor performance in many-arms scenario, and we think that 10000 samples, although they are sufficiently many for 5 arms case, are not sufficient enough for 10 and 20 arms case, as there are more Bayesian estimates for TS to learn as the number of arms grow. All algorithms tested have regret bounds for Gaussian MAB problems tighter than the bound of ORKG we present in Theorem 1, and this empirical validation suggests existence of tighter regret bounds for ORKG.

6 Discussion

The simple regularization method for KG used in ORKG algorithm allows the first KG-based algorithm with sublinear regret bounds, yet this approach may

be too simple to tighten regret bounds of ORKG on par with other stochastic MAB algorithms. Despite the theoretical gap in regret bounds, we witness impressive empirical performance of ORKG in MAB benchmarks with correct model specification. Notably, the empirical validations is performed with relatively few samples from MAB perspective, which suggests ORKG can perform well in real world applications where the number of samples are limited. Also, ORKG gives new insight to a long-standing question in KG literature on how to trade off exploration-exploitation correctly in online learning, and at the same time, ORKG allows interdisciplinary discussion between KG and MAB literature by providing the first regret bound result of KG-based algorithm in MAB problems.

7 Conclusion

We present a simple and effective method to regularize knowledge gradient (KG) that allows novel asymptotic regret analysis of KG-based algorithms with independent Gaussian belief model. Using regularized knowledge gradients, we construct ORKG, a KG-based online learning algorithm, and present its sublinear regret bound in partial information Gaussian MAB problem. We provide empirical validation of ORKG, and verify that ORKG algorithm performs comparable to select MAB algorithms with tighter regret bounds in Gaussian MAB benchmarks. Our result opens up an interesting stage for further research in KG from the perspective of MAB literature.

References

1. Abramowitz, M., Stegun, I.A. (eds.): Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,. Dover Publications, Inc., New York, NY, USA (1974)
2. Auer, P., Cesa-Bianchi, N., Freund, Y., Schapire, R.E.: The Nonstochastic Multi-armed Bandit Problem. SIAM Journal on Computing **32**(1), 48–77 (Jan 2002)
3. Besson, L.: SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python. Online at: GitHub.com/SMPyBandits/SMPyBandits (2018)
4. Cesa-Bianchi, N., Gentile, C., Lugosi, G., Neu, G.: Boltzmann Exploration Done Right. Advances in neural information processing systems **30** (2017)
5. Chen, S., Reyes, K.R.G., Gupta, M.K., McAlpine, M.C., Powell, W.B.: Optimal learning in experimental design using the knowledge gradient policy with application to characterizing nanoemulsion stability. SIAM/ASA Journal on Uncertainty Quantification **3**(1), 320–345 (2015)
6. Frazier, P., Powell, W.: The Knowledge Gradient Policy for Offline Learning with Independent Normal Rewards (2007)
7. Frazier, P., Powell, W., Dayanik, S.: The Knowledge-Gradient Policy for Correlated Normal Beliefs. INFORMS journal on computing **21**(4), 599–613 (Nov 2009)
8. Frazier, P.I., Powell, W.B., Dayanik, S.: A Knowledge-Gradient Policy for Sequential Information Collection. SIAM Journal on Control and Optimization **47**(5), 2410–2439 (Jan 2008)

9. Garivier, A., Cappé, O.: The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond. In: Kakade, S.M., von Luxburg, U. (eds.) Proceedings of the 24th Annual Conference on Learning Theory. Proceedings of Machine Learning Research, vol. 19, pp. 359–376. PMLR, Budapest, Hungary (2011)
10. Han, W., Powell, W.B.: Optimal Online Learning for Nonlinear Belief Models Using Discrete Priors. Operations research **68**(5), 1538–1556 (Sep 2020)
11. He, X., Reyes, K.G., Powell, W.B.: Optimal Learning with Local Nonlinear Parametric Models over Continuous Designs. SIAM Journal of Scientific Computing **42**(4), A2134–A2157 (Jan 2020)
12. Huang, Y., Zhao, L., Powell, W.B., Tong, Y., Ryzhov, I.O.: Optimal Learning for Urban Delivery Fleet Allocation. Transportation Science **53**(3), 623–641 (May 2019)
13. Lai, T.L., Robbins, H.: Asymptotically efficient adaptive allocation rules. Advances in applied mathematics **6**(1), 4–22 (1985)
14. Negoescu, D.M., Frazier, P.I., Powell, W.B.: The Knowledge-Gradient Algorithm for Sequencing Experiments in Drug Discovery (2011)
15. Ryzhov, I.O., Powell, W.: The knowledge gradient algorithm for online subset selection. In: 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning. pp. 137–144 (Mar 2009)
16. Ryzhov, I.O., Powell, W.B., Frazier, P.I.: The Knowledge Gradient Algorithm for a General Class of Online Learning Problems. Operations research **60**(1), 180–195 (Feb 2012)
17. Scott, W., Frazier, P., Powell, W.: The Correlated Knowledge Gradient for Simulation Optimization of Continuous Parameters using Gaussian Process Regression. SIAM journal on optimization: a publication of the Society for Industrial and Applied Mathematics **21**(3), 996–1026 (Jul 2011)
18. Seldin, Y., Slivkins, A.: One Practical Algorithm for Both Stochastic and Adversarial Bandits. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 1287–1295. PMLR, Bejing, China (2014)
19. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.: Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design (Dec 2009)
20. Thompson, W.R.: On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. Biometrika **25**(3/4), 285–294 (1933)
21. Thul, L., Powell, W.: Stochastic Optimization for Vaccine and Testing Kit Allocation for the COVID-19 Pandemic (Jan 2021)
22. Tian, Z., Han, W., Powell, W.B.: Adaptive Learning of Drug Quality and Optimization of Patient Recruitment for Clinical Trials with Dropouts. Manufacturing & Service Operations Management (Mar 2021)
23. Wang, Y., Do Nascimento, J.M., Powell, W.: Reinforcement Learning for Dynamic Bidding in Truckload Markets: an Application to Large-Scale Fleet Management with Advance Commitments (Feb 2018)
24. Wang, Y., Wang, C., Powell, W.: The Knowledge Gradient for Sequential Decision Making with Stochastic Binary Feedbacks. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1138–1147. PMLR, New York, New York, USA (2016)