Regularized Knowledge Gradient

Donghun Lee

Dept. of Math
Korea Univ.

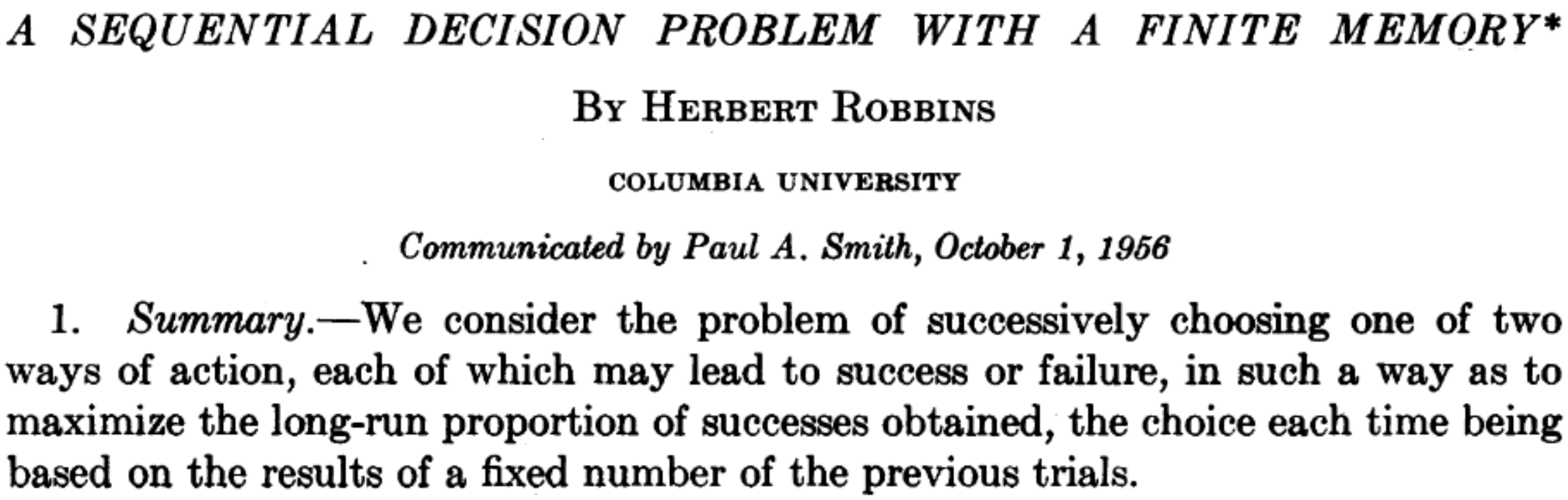
1 November, 2022

AIML@K

KOREA UNIVERSITY

Problem Setting – Historical

- Sequential Decision Making
  - Relatively new problem?



*A SEQUENTIAL DECISION PROBLEM WITH A FINITE MEMORY**

BY HERBERT ROBBINS

COLUMBIA UNIVERSITY

Communicated by Paul A. Smith, October 1, 1956

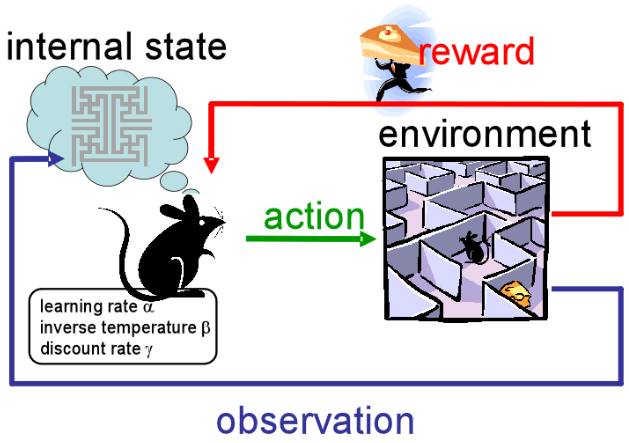
1. *Summary.*—We consider the problem of successively choosing one of two ways of action, each of which may lead to success or failure, in such a way as to maximize the long-run proportion of successes obtained, the choice each time being based on the results of a fixed number of the previous trials.

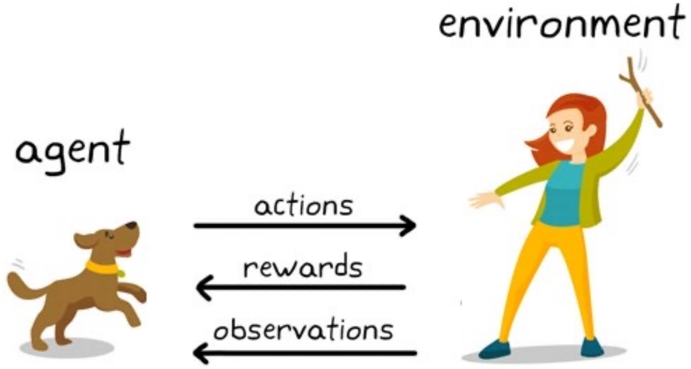
  - Actually, it's a problem of the entire (human?) history

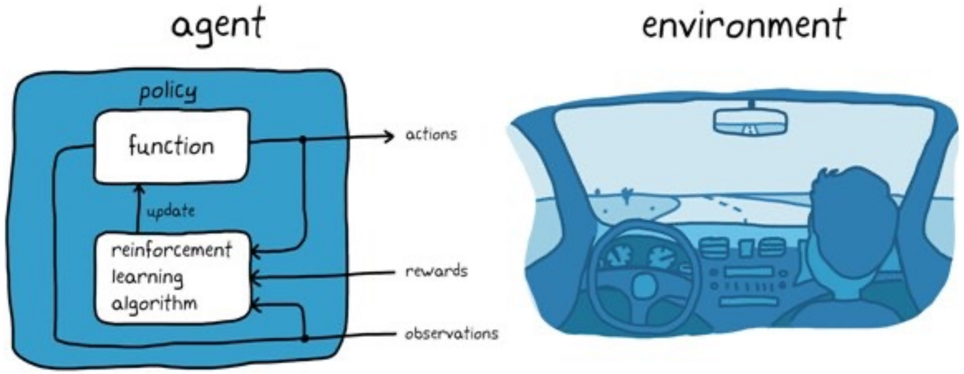
Problem Setting – Modern

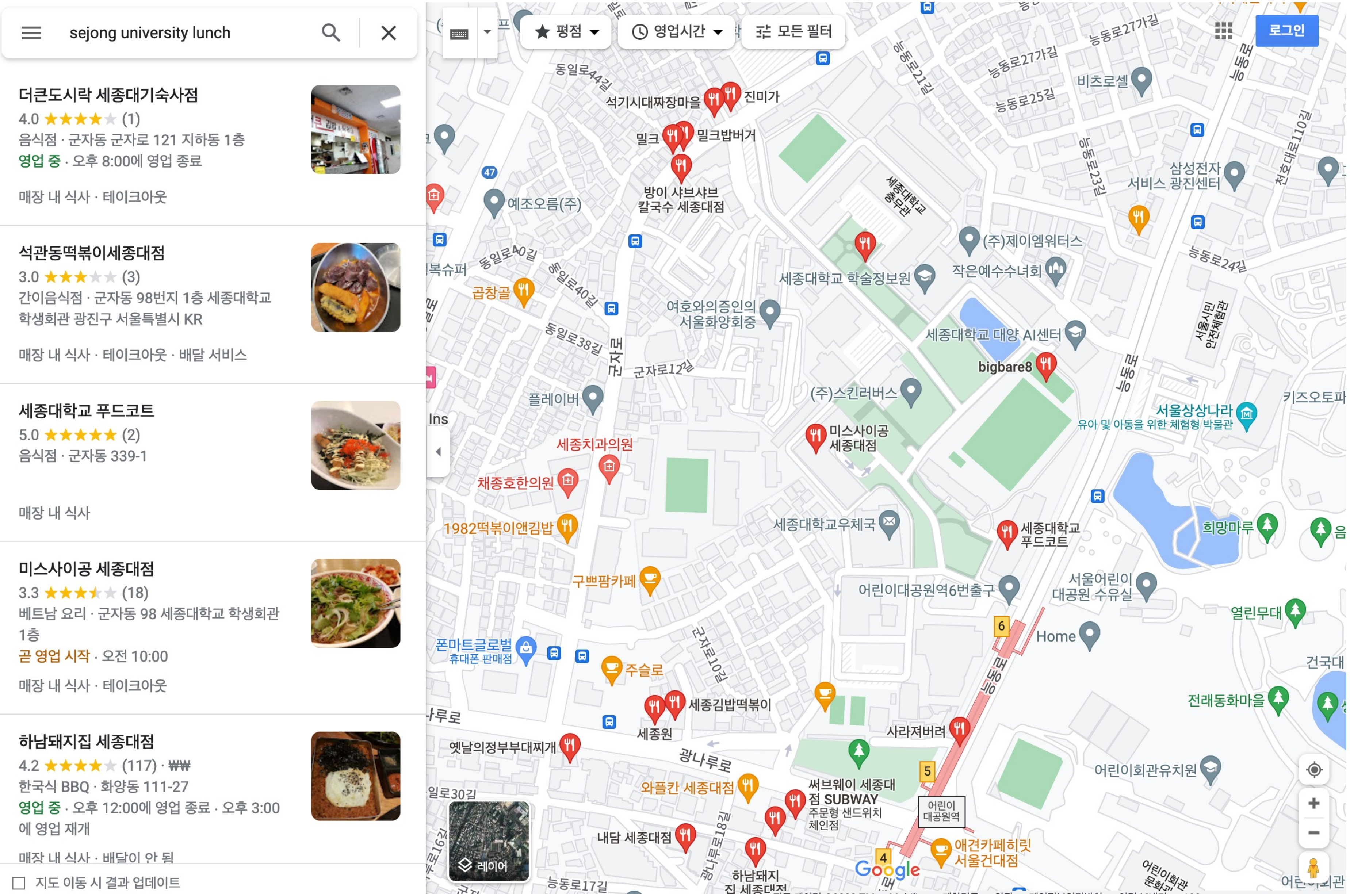
- Sequential Decision Making
  - Sounds similar to …
    - Reinforcement Learning
    - Control Theory

- How to make “best” decision?
  - Now, and also later
  - Based on finite interactions
  - With the aim of optimizing some fn.









sejong university lunch

더큰도시락 세종대기숙사점
4.0 ★★★★☆ (1)
음식점 · 군자동 군자로 121 지하동 1층
영업 중 · 오후 8:00에 영업 종료
매장 내 식사 · 테이크아웃

석관동떡볶이세종대점
3.0 ★★★☆☆ (3)
간이음식점 · 군자동 98번지 1층 세종대학교 학생회관 광진구 서울특별시 KR
매장 내 식사 · 테이크아웃 · 배달 서비스

세종대학교 푸드코트
5.0 ★★★★★ (2)
음식점 · 군자동 339-1
매장 내 식사

미스사이공 세종대점
3.3 ★★★☆☆ (18)
베트남 요리 · 군자동 98 세종대학교 학생회관 1층
곧 영업 시작 · 오전 10:00
매장 내 식사 · 테이크아웃

하남돼지집 세종대점
4.2 ★★★★☆ (117) · ₩₩
한국식 BBQ · 화양동 111-27
영업 중 · 오후 12:00에 영업 종료 · 오후 3:00에 영업 재개
매장 내 식사 · 배달이 안 됨

☐ 지도 이동 시 결과 업데이트

Problem Setting – Everyday

- Decision: what to eat for lunch on n-th day
  - E.g. represented by a finite set $\mathcal{X} = \{0,1,2,3,4\}$

$x^n = 0$

$x^n = 1$

$x^n = 2$

$x^n = 3$

$x^n = 4$

Problem Setting – Everyday

- Reward / response: $R(x)$
  - "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)

$R(0)$

$R(1)$

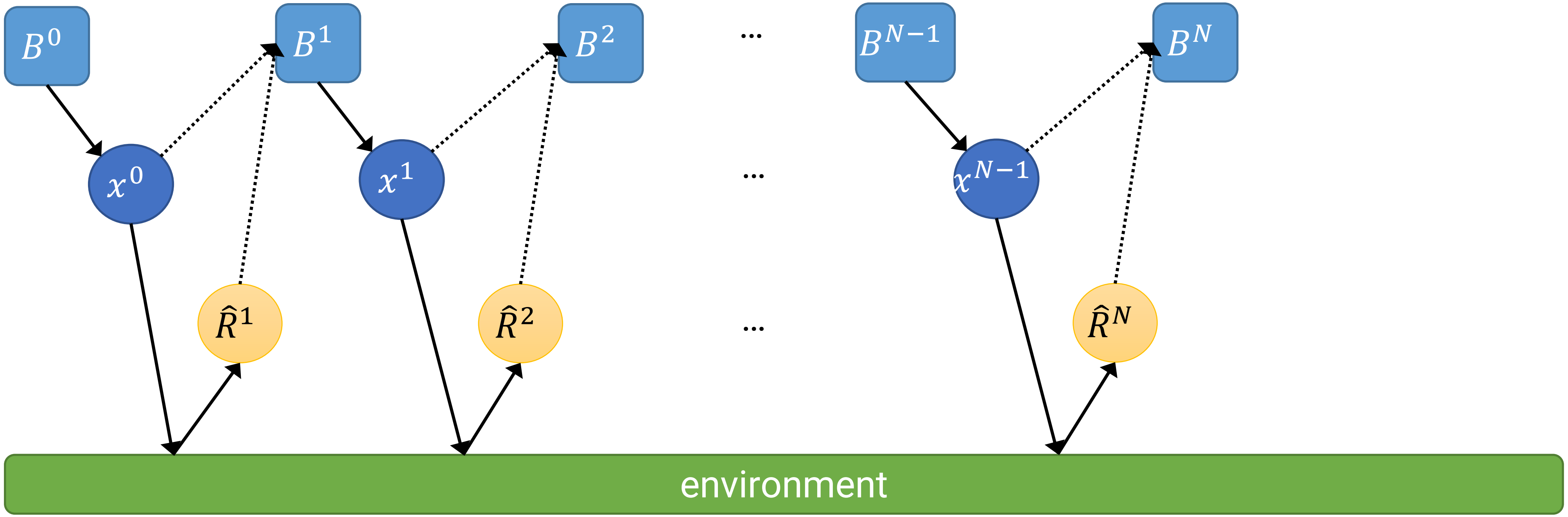
$R(2)$

$R(3)$

$R(4)$

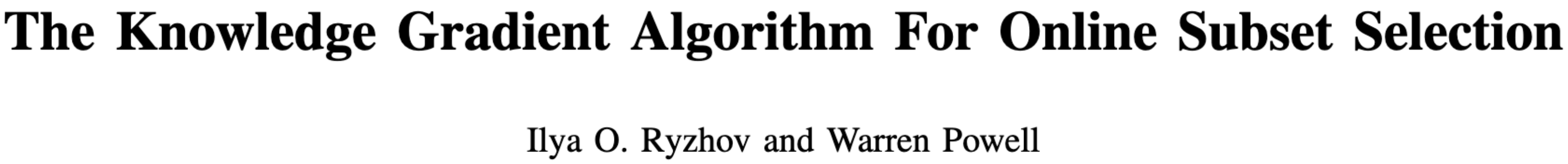
Canonical Problem Formulation

- Decision: finite set $\mathcal{X}$
- Time horizon / budget: N



Knowledge Gradient in Online Problems

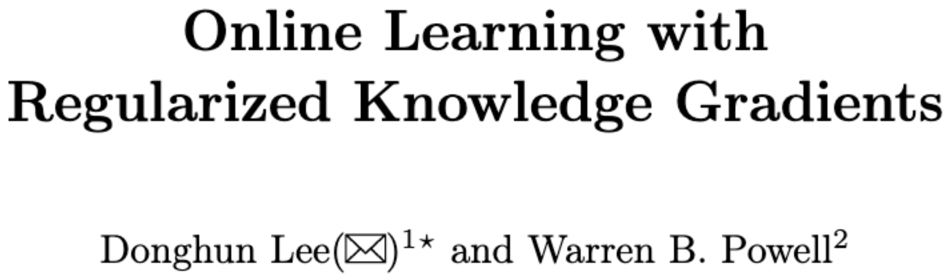
- Ryzhov and Powell, 2009



The Knowledge Gradient Algorithm For Online Subset Selection

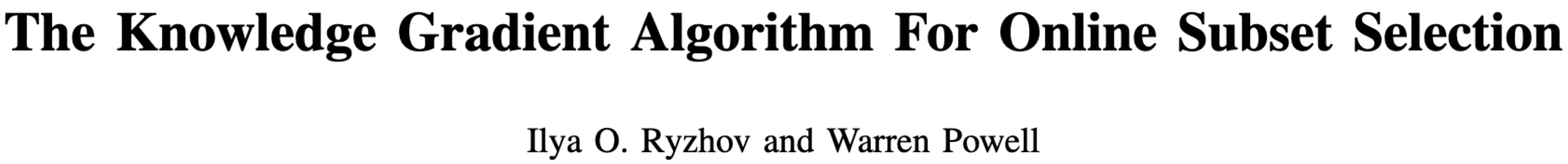
Ilya O. Ryzhov and Warren Powell

- Lee and Powell, 2022



Online Learning with Regularized Knowledge Gradients

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



The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

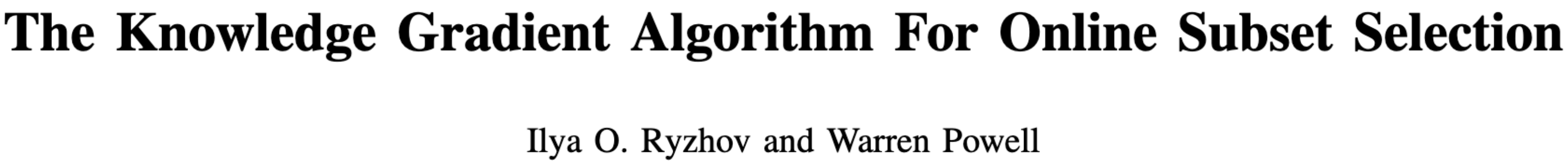
Ryzhov and Powell, 2009

Problem Definition

- (Admissible) Subset Selection Problem
  - Given a finite set $\mathcal{U}$, find $\mathcal{X} \subseteq \mathcal{U}$ s.t.$\forall i \in \{1, \cdots, k\}$: $c_i = T$
    - Where $\forall i \in \{1, \cdots, k\}, c_i\colon pow(\mathcal{U}) \to \{T, F\}$
      - “constraints”

Problem Definition

- (Admissible) Subset Selection Problem
  - Given a finite set $\mathcal{U}$, find $\mathcal{X} \subseteq \mathcal{U}$ s.t.$\forall i \in \{1, \cdots, k\}$: $c_i = T$
    - Where $\forall i \in \{1, \cdots, k\}, c_i : pow(\mathcal{U}) \rightarrow \{T, F\}$
      - "constraints"

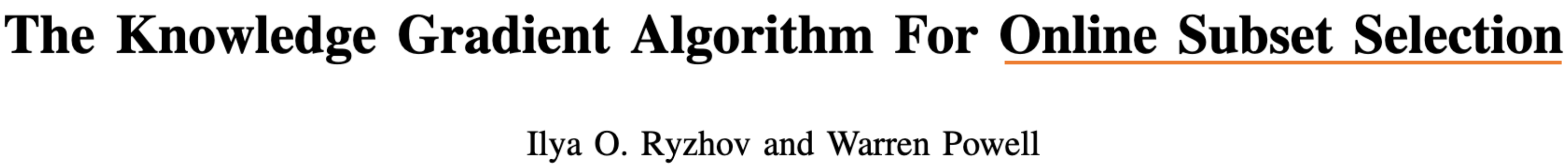


The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

- (Online) "Subset Selection Problem"
  - Given a finite set of choices and a number N, find "best" length-N-sequence of choices

Problem Definition

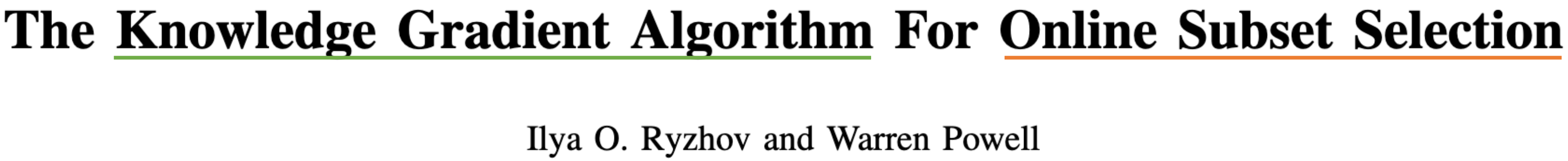


The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

- (Online) “Subset Selection Problem”
  - Given a finite set of choices and a number N, find ”best” length-N-sequence of choices
  - Free variables: allocate N measurements
  - Objective: maximize the sum of rewards from all N measurements

Problem Definition



The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

- (Online) “Subset Selection Problem”
  - Given a finite set of choices and a number N, find ”best” length-N-sequence of choices

  - Free variables: allocate N measurements
  - Objective: maximize the sum of rewards from all N measurements

- Algorithm that uses KG

Problem Definition

The Knowledge Gradient Algorithm For Online Subset Selection

Ilya O. Ryzhov and Warren Powell

- Multi-armed Bandit Problem
  - Given a finite set of **arms** and a **finite horizon** N, find ”best” length-N-sequence of choices

  - Free variables: allocate N measurements
  - Objective: maximize the sum of rewards from all N measurements

- Algorithm that uses KG

Online KG Policy

- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \operatorname*{argmax}_{x \in \mathcal{X}} \mu_x^n + (N - n)\nu_x^{KG,n}$$

Online KG Policy

- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \operatorname*{argmax}_{x \in \mathcal{X}} \mu_x^n + (N - n)\nu_x^{KG,n}$$

- Key idea: given state $s^n = \{(\mu_x^n, \sigma_x^n)\}_{x \in \mathcal{X}}$ (at iteration n)
  - **Expected** reward (at n) of choosing $x \in \mathcal{X}$: μ_x^n
  - **Expected** reward (at n) of choosing "best" x (w/ belief at n): $\max_{x \in \mathcal{X}} \mu_x^n$

  - **Expected** total reward (from n till end of horizon) of choosing "best" x (w/ belief at n): $(N - n + 1)\max_{x \in \mathcal{X}} \mu_x^n$
    - Let $V^{EB,n}(s^n) \coloneqq (N - n + 1)\max_{x \in \mathcal{X}} \mu_x^n$ ("Empirical Bayesian" policy)

Online KG Policy

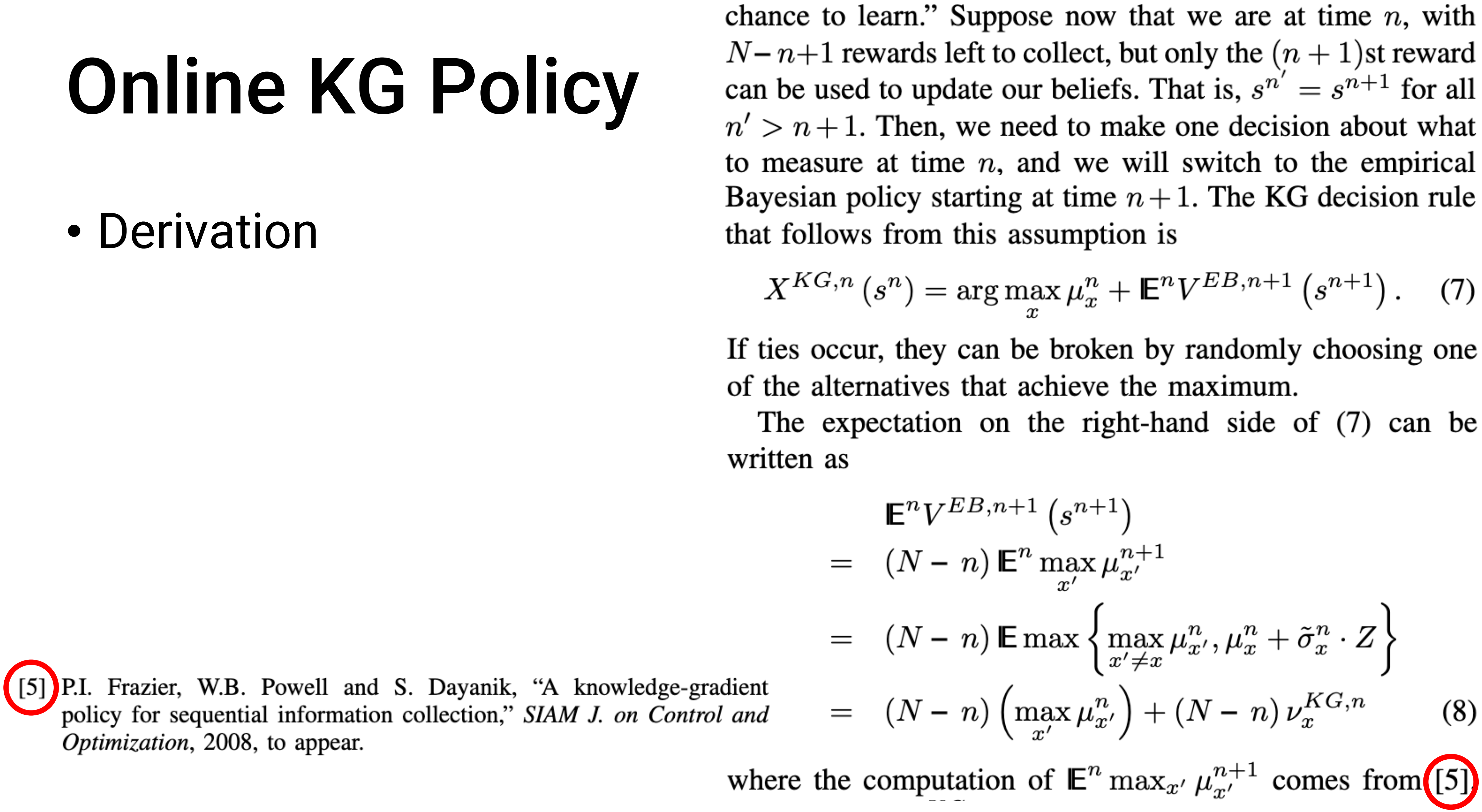
- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, \mu_x^n + (N-n)\nu_x^{KG,n}$$

- Key idea: given state $s^n = \{(\mu_x^n, \sigma_x^n)\}_{x \in \mathcal{X}}$ (at iteration n)
  - **Expected** reward (at n) of choosing $x \in \mathcal{X}$: μ_x^n
  - **Expected** reward (at n) of choosing "best" x (w/ belief at n): $\max_{x \in \mathcal{X}} \mu_x^n$
  - **Expected** total reward (from n till end of horizon) of choosing "best" x (w/ belief at n): $(N-n+1)\max_{x \in \mathcal{X}} \mu_x^n$
    - Let $V^{EB,n}(s^n) \coloneqq (N-n+1)\max_{x \in \mathcal{X}} \mu_x^n$ ("Empirical Bayesian" policy)

Online KG Policy

- Derivation



chance to learn." Suppose now that we are at time n, with $N-n+1$ rewards left to collect, but only the $(n+1)$st reward can be used to update our beliefs. That is, $s^{n'} = s^{n+1}$ for all $n' > n+1$. Then, we need to make one decision about what to measure at time n, and we will switch to the empirical Bayesian policy starting at time $n+1$. The KG decision rule that follows from this assumption is

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu_x^n + \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right). \qquad (7)$$

If ties occur, they can be broken by randomly choosing one of the alternatives that achieve the maximum.

The expectation on the right-hand side of (7) can be written as

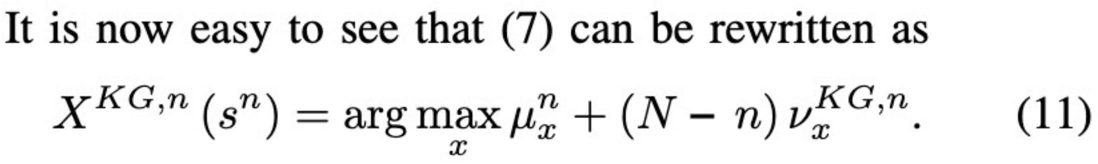
$$\begin{aligned} & \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right) \\ = \; & (N-n)\, \mathbb{E}^n \max_{x'} \mu_{x'}^{n+1} \\ = \; & (N-n)\, \mathbb{E} \max\left\{ \max_{x' \neq x} \mu_{x'}^n, \mu_x^n + \tilde{\sigma}_x^n \cdot Z \right\} \\ = \; & (N-n)\left(\max_{x'} \mu_{x'}^n\right) + (N-n)\, \nu_x^{KG,n} \qquad (8) \end{aligned}$$

where the computation of $\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1}$ comes from [5]

[5] P.I. Frazier, W.B. Powell and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM J. on Control and Optimization*, 2008, to appear.

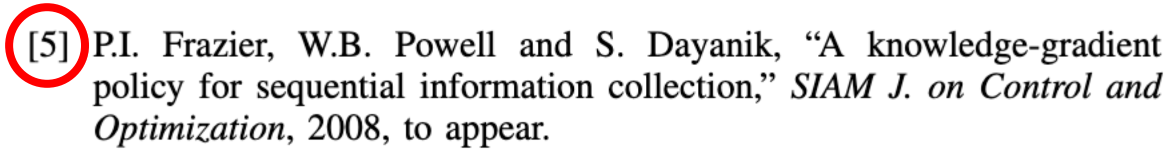
Online KG Policy

- Derivation
  - Substitute (8) to (7)
  - Remove constants

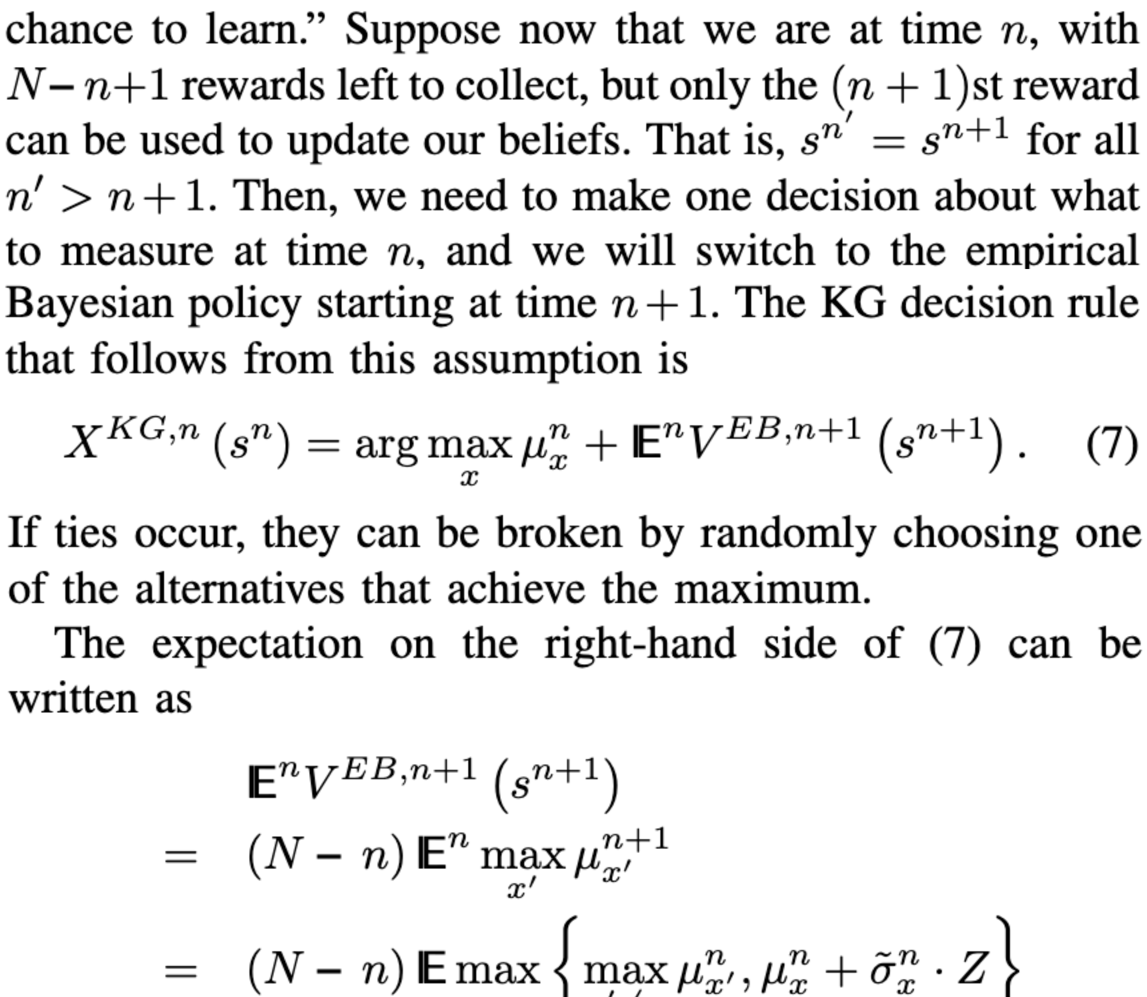


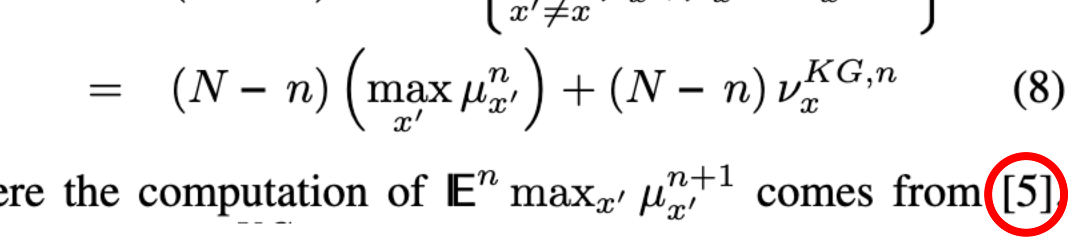
It is now easy to see that (7) can be rewritten as

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu_x^n + (N-n)\,\nu_x^{KG,n}. \qquad (11)$$



[5] P.I. Frazier, W.B. Powell and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM J. on Control and Optimization*, 2008, to appear.





chance to learn." Suppose now that we are at time n, with $N-n+1$ rewards left to collect, but only the $(n+1)$st reward can be used to update our beliefs. That is, $s^{n'} = s^{n+1}$ for all $n' > n+1$. Then, we need to make one decision about what to measure at time n, and we will switch to the empirical Bayesian policy starting at time $n+1$. The KG decision rule that follows from this assumption is

$$X^{KG,n}\left(s^n\right) = \arg\max_x \mu_x^n + \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right). \qquad (7)$$

If ties occur, they can be broken by randomly choosing one of the alternatives that achieve the maximum.

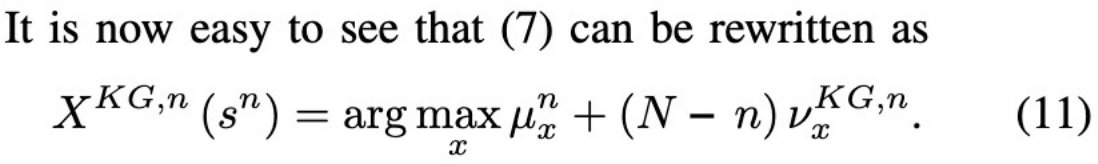
The expectation on the right-hand side of (7) can be written as

$$\begin{aligned}
& \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right) \\
= \; & (N-n)\,\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1} \\
= \; & (N-n)\,\mathbb{E}\max\left\{\max_{x'\neq x} \mu_{x'}^n, \mu_x^n + \tilde{\sigma}_x^n \cdot Z\right\} \\
= \; & (N-n)\left(\max_{x'} \mu_{x'}^n\right) + (N-n)\,\nu_x^{KG,n} \qquad (8)
\end{aligned}$$

where the computation of $\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1}$ comes from [5]

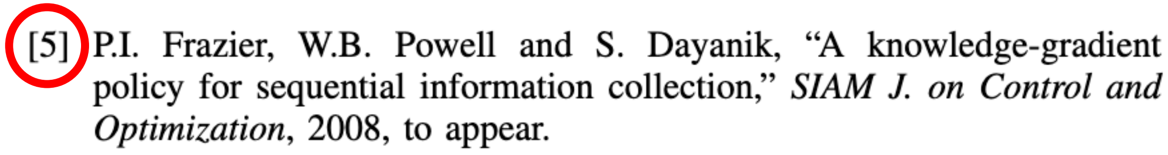
Online KG Policy

- Derivation
    - Substitute (8) to (7)
    - Remove constants

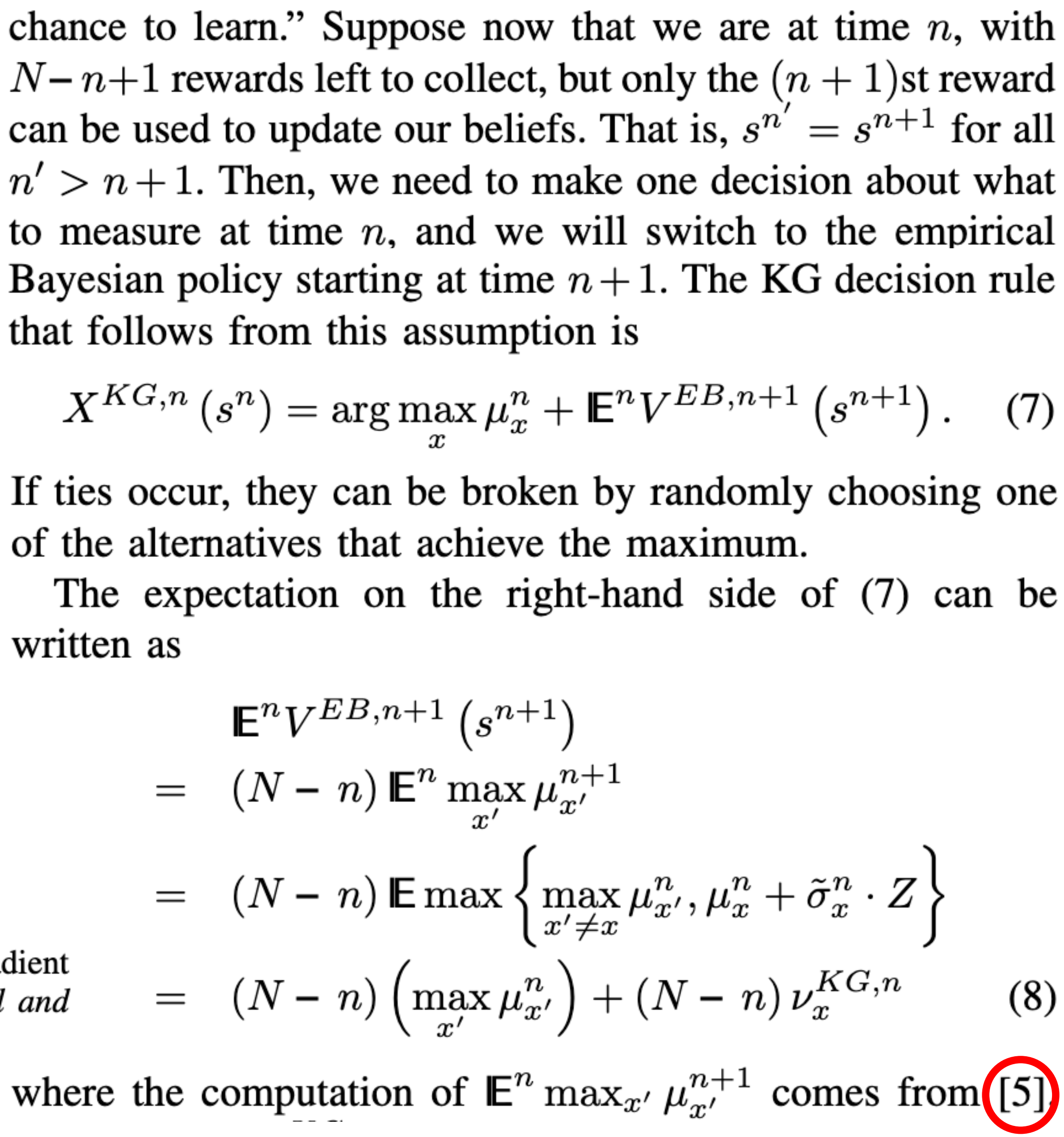


It is now easy to see that (7) can be rewritten as

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + (N-n)\,\nu_x^{KG,n}. \qquad (11)$$



[5] P.I. Frazier, W.B. Powell and S. Dayanik, "A knowledge-gradient policy for sequential information collection," *SIAM J. on Control and Optimization*, 2008, to appear.



chance to learn." Suppose now that we are at time n, with $N-n+1$ rewards left to collect, but only the $(n+1)$st reward can be used to update our beliefs. That is, $s^{n'} = s^{n+1}$ for all $n' > n+1$. Then, we need to make one decision about what to measure at time n, and we will switch to the empirical Bayesian policy starting at time $n+1$. The KG decision rule that follows from this assumption is

$$X^{KG,n}(s^n) = \arg\max_x \mu_x^n + \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right). \qquad (7)$$

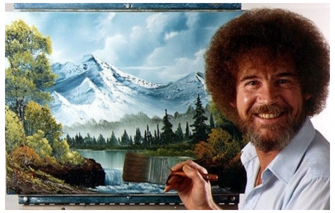
If ties occur, they can be broken by randomly choosing one of the alternatives that achieve the maximum.

The expectation on the right-hand side of (7) can be written as

$$\begin{aligned} & \mathbb{E}^n V^{EB,n+1}\left(s^{n+1}\right) \\ = \; & (N-n)\,\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1} \\ = \; & (N-n)\,\mathbb{E}\max\left\{\max_{x'\neq x} \mu_{x'}^n, \mu_x^n + \tilde{\sigma}_x^n \cdot Z\right\} \\ = \; & (N-n)\left(\max_{x'} \mu_{x'}^n\right) + (N-n)\,\nu_x^{KG,n} \qquad (8) \end{aligned}$$

where the computation of $\mathbb{E}^n \max_{x'} \mu_{x'}^{n+1}$ comes from [5]

Online KG Policy, Revisited



Ain't it easy..?!!

- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, \mu_x^n + (N-n)\nu_x^{KG,n}$$

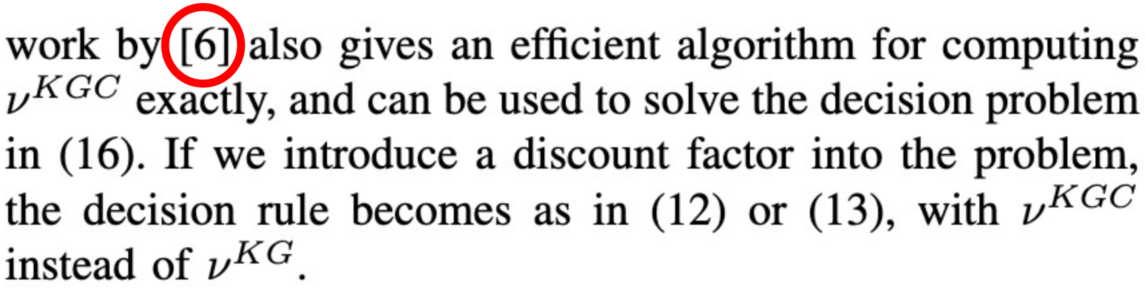
- Key idea: given state $s^n = \{(\mu_x^n, \sigma_x^n)\}_{x \in \mathcal{X}}$ (at iteration n)
  - **Expected** reward (at n) of choosing $x \in \mathcal{X}$: μ_x^n
  - **Expected** reward (at n) of choosing "best" x (w/ belief at n): $\max_{x \in \mathcal{X}} \mu_x^n$
  - **Expected** total reward (from n till end of horizon) of choosing "best" x (w/ belief at n): $(N-n+1)\max_{x \in \mathcal{X}} \mu_x^n$
    - Let $V^{EB,n}(s^n) \coloneqq (N-n+1)\max_{x \in \mathcal{X}} \mu_x^n$ ("Empirical Bayesian" policy)

Online KG Policy, Correlated Belief

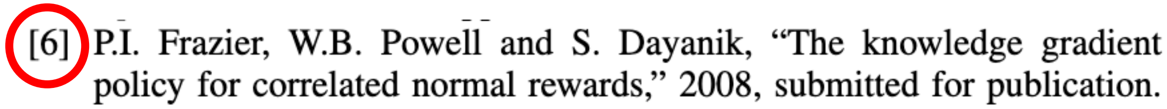
- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \, \mu_x^n + (N-n)\nu_x^{KG,n}$$

- Substitute update rule for s^n from independent belief to correlated belief



work by [6] also gives an efficient algorithm for computing ν^{KGC} exactly, and can be used to solve the decision problem in (16). If we introduce a discount factor into the problem, the decision rule becomes as in (12) or (13), with ν^{KGC} instead of ν^{KG}.



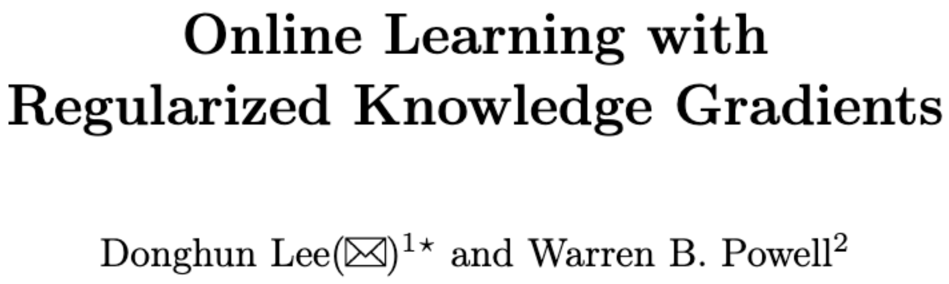
[6] P.I. Frazier, W.B. Powell and S. Dayanik, "The knowledge gradient policy for correlated normal rewards," 2008, submitted for publication.

Ain't it easy..?!!

Can’t We Do Better?

- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \, \mu_x^n + (N - n)\nu_x^{KG,n}$$



Online Learning with
Regularized Knowledge Gradients

Donghun Lee(✉)^{1*} and Warren B. Powell^{2}

Lee and Powell, 2022

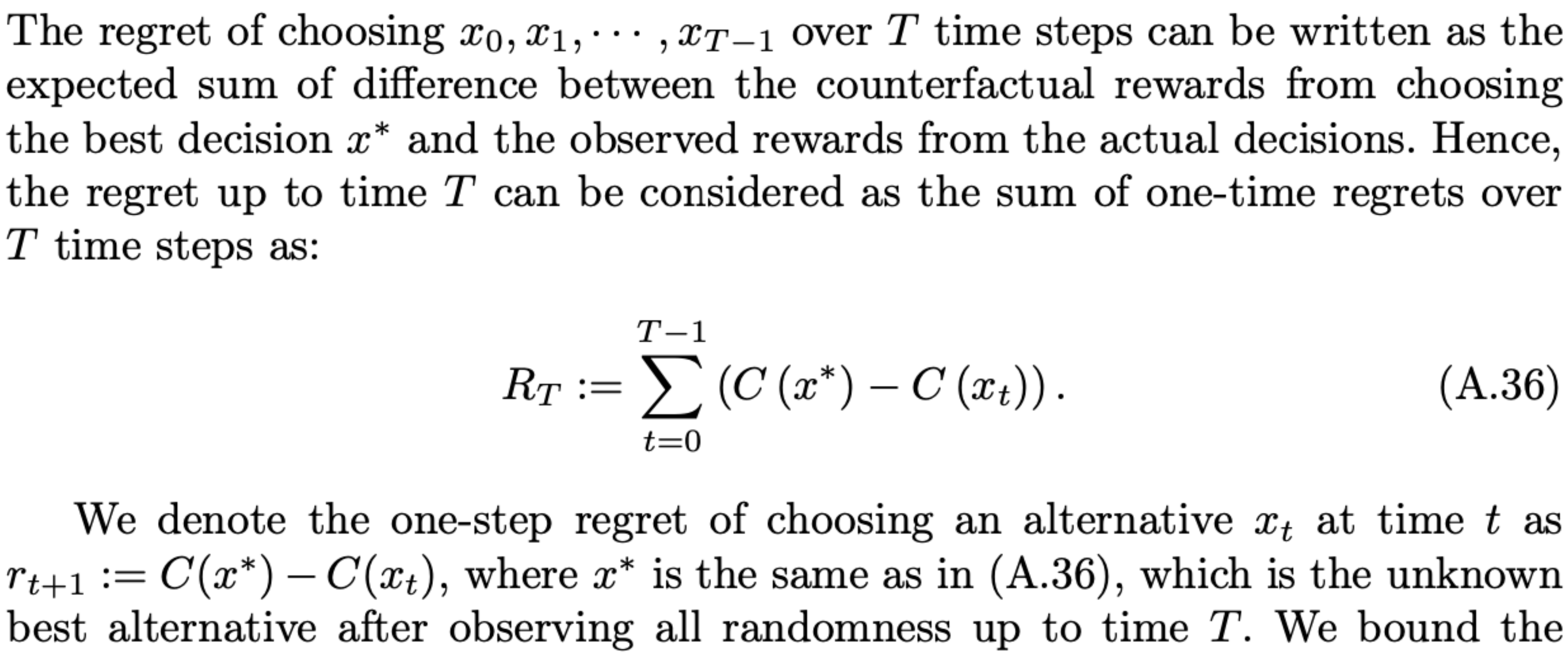
Most work done in 2019-2020

Reward and Regret

- Reward (of choosing action x given belief B_t)
  - True reward $\forall t, \forall x \in \mathcal{X}$: $R(x) \sim N(\mu^x, (\sigma^x)^2)$
    - Stationary stochastic MAB, Gaussian arms
  - Surrogate reward $\forall t, \forall x \in \mathcal{X}$: $R(B_t, x) \sim N(\bar{\mu}_t^x, (\bar{\sigma}_t^x)^2)$
    - Based on current belief B_t

Reward and Regret

- Regret defined as ”sum of missed rewards”



The regret of choosing $x_0, x_1, \cdots, x_{T-1}$ over T time steps can be written as the expected sum of difference between the counterfactual rewards from choosing the best decision x^* and the observed rewards from the actual decisions. Hence, the regret up to time T can be considered as the sum of one-time regrets over T time steps as:

$$R_T := \sum_{t=0}^{T-1} \left(C\left(x^*\right) - C\left(x_t\right)\right). \tag{A.36}$$

We denote the one-step regret of choosing an alternative x_t at time t as $r_{t+1} := C(x^*) - C(x_t)$, where x^* is the same as in (A.36), which is the unknown best alternative after observing all randomness up to time T. We bound the

Online Regularized KG (ORKG)

- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \left\{ \mu_x^n + (N - n) \nu_x^{KG,n} \right\}$$

- Online Regularized KG policy (Lee and Powell 2022):

$$\pi^{ORKG}(s^n) = \underset{x \in \mathcal{X}}{\operatorname{argmax}} \left\{ \mu_x^n + \rho(n) \cdot \nu_x^{RKG,n} \right\}$$

Online Regularized KG (ORKG)

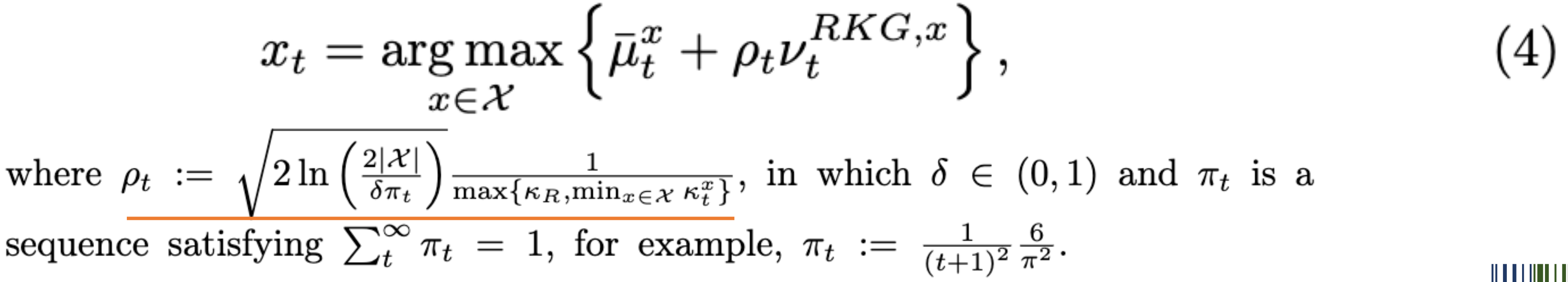
- Online KG policy (Ryzhov and Powell 2009):

$$\pi^{OKG}(s^n) = \underset{x\in\mathcal{X}}{\text{argmax}}\{\mu_x^n + (N-n)\nu_x^{KG,n}\}$$

- Online Regularized KG policy (Lee and Powell 2022):

$$\pi^{ORKG}(s^n) = \underset{x\in\mathcal{X}}{\text{argmax}}\{\mu_x^n + \rho(n)\cdot\nu_x^{RKG,n}\}$$

  - i.e.

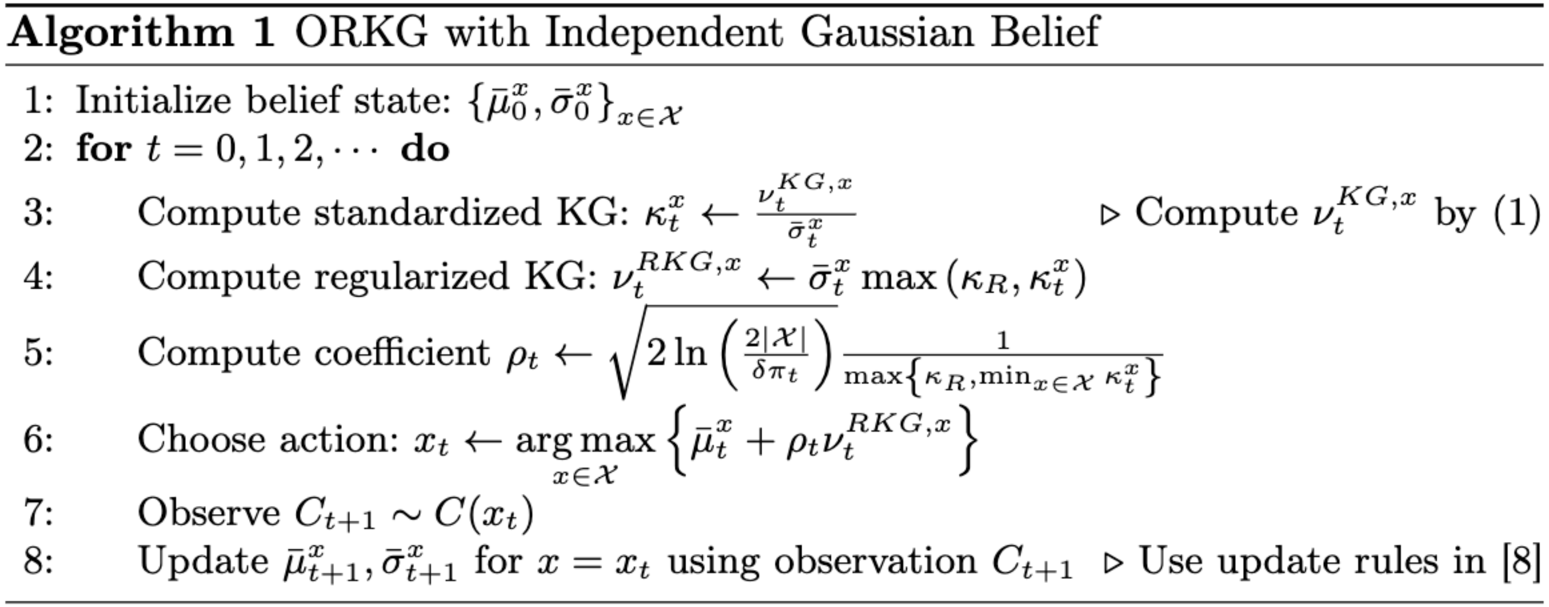


$$x_t = \arg\max_{x\in\mathcal{X}}\left\{\bar{\mu}_t^x + \rho_t\nu_t^{RKG,x}\right\}, \tag{4}$$

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, $\pi_t := \frac{1}{(t+1)^2}\frac{6}{\pi^2}$.

Online Regularized KG (ORKG)

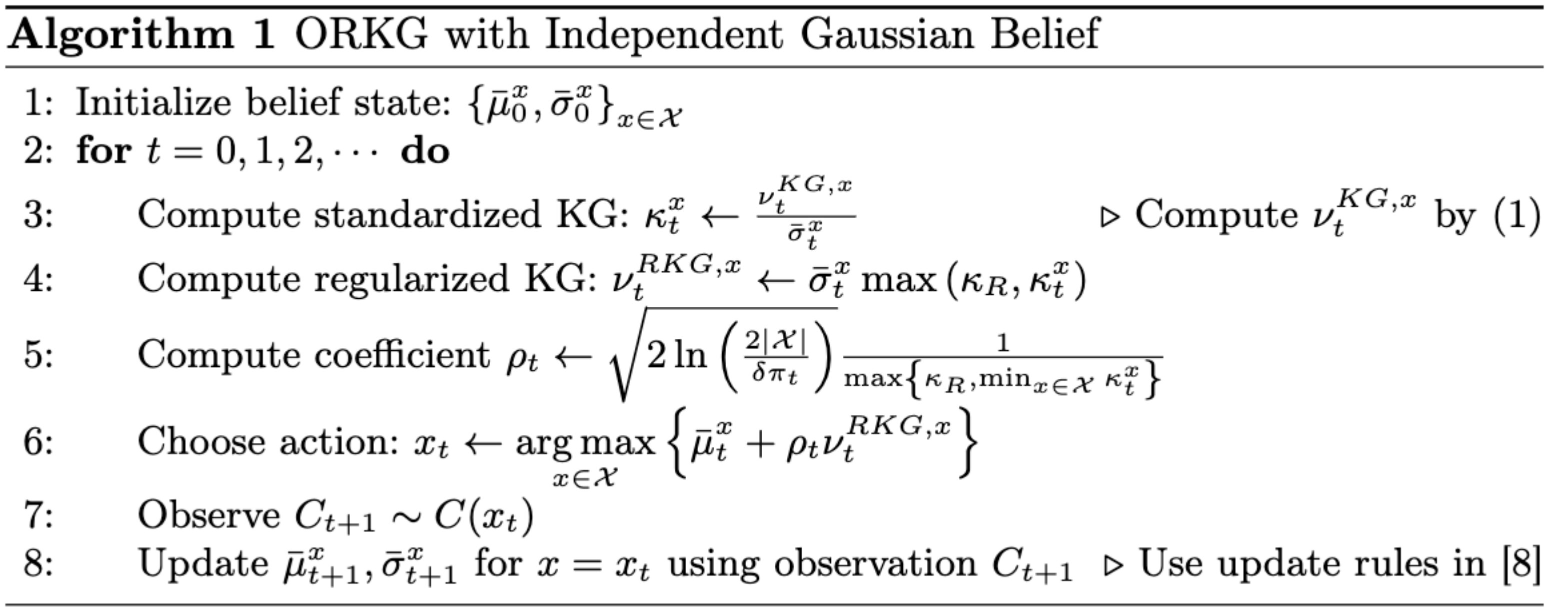
- Algorithm for online learning
  - Using regularized KG



Algorithm 1 ORKG with Independent Gaussian Belief

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

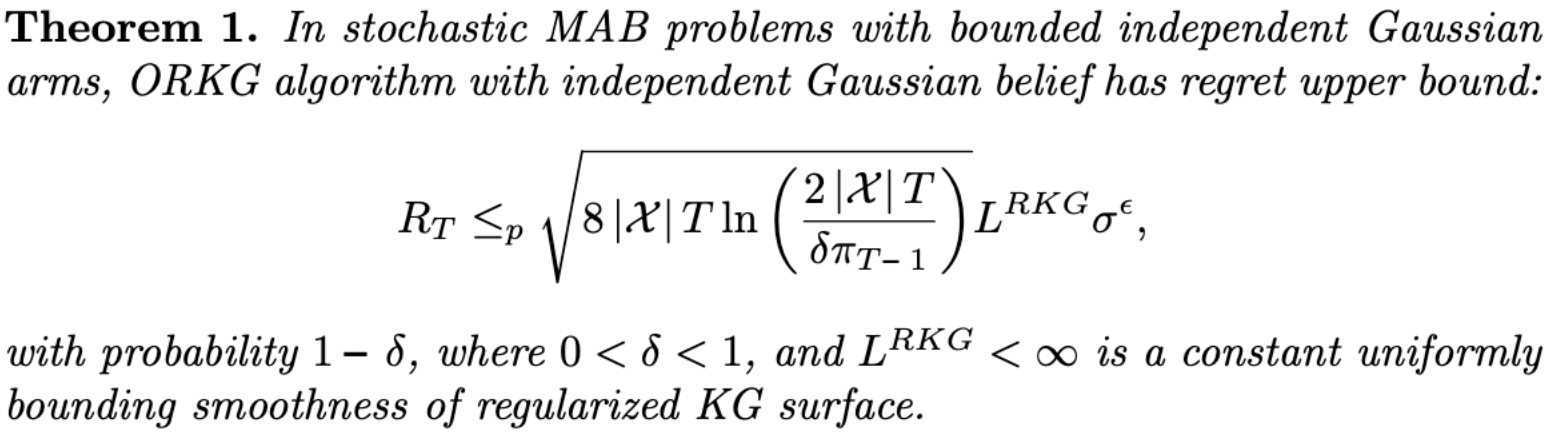
ORKG



Algorithm 1 ORKG with Independent Gaussian Belief

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

- … that has sublinear regret bound (first in KG algorithms)



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

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

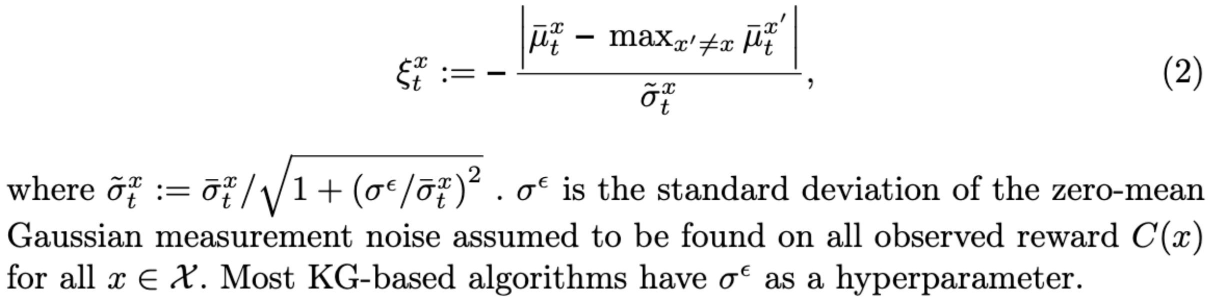
with probability $1 - \delta$, where $0 < \delta < 1$, and $L^{RKG} < \infty$ is a constant uniformly bounding smoothness of regularized KG surface.

How?

- Add suitable properties to KG
  - Standardize KG
  - Regularize KG

- Get probabilistic bounds given KG-based choices
  - Bound one-step reward deviation w/ high probability
  - Bound one-step regret
  - Bound sum of regrets over T-steps

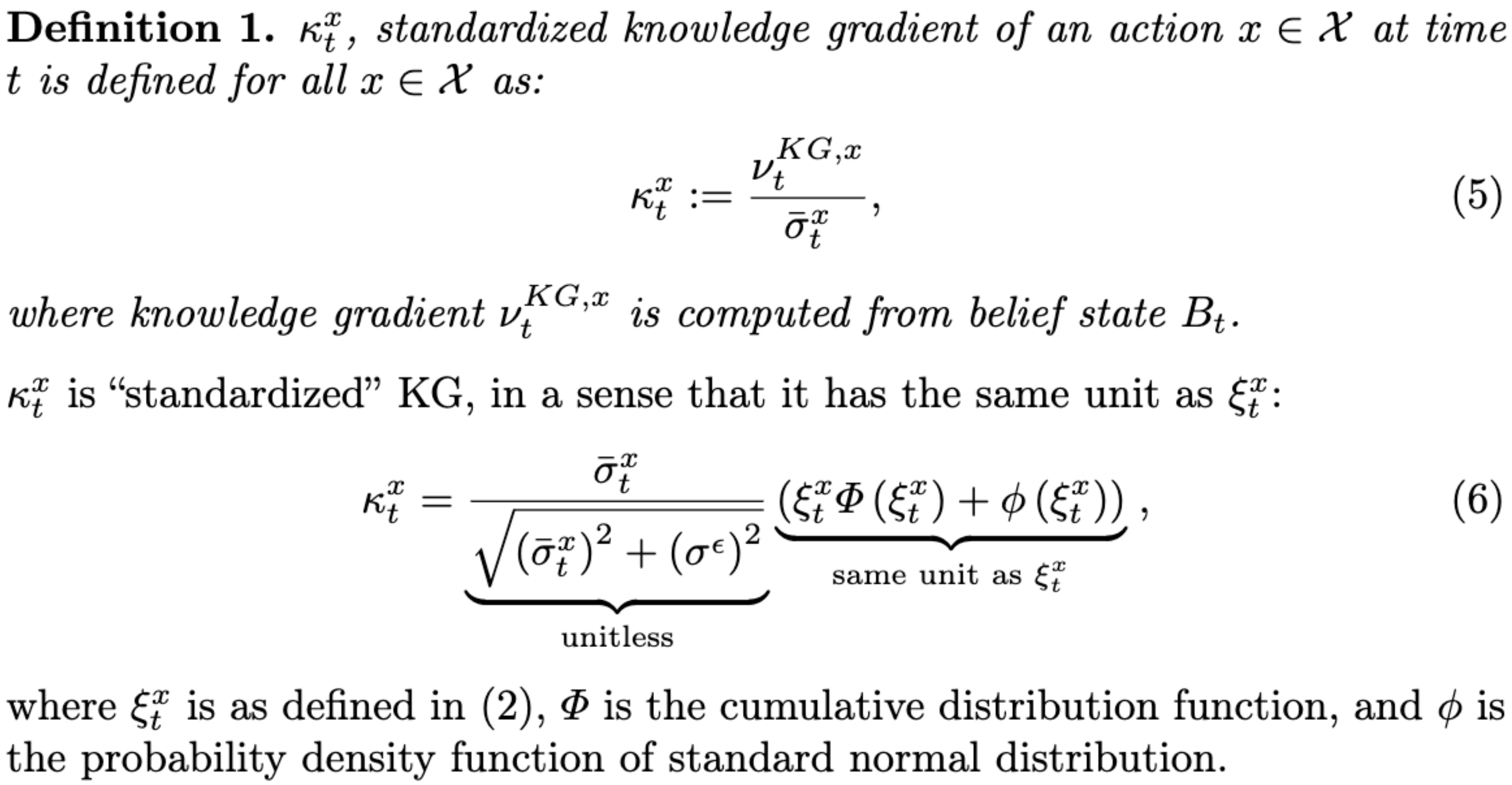
Prep Work



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

where $\tilde{\sigma}_t^x := \bar{\sigma}_t^x / \sqrt{1 + (\sigma^\epsilon / \bar{\sigma}_t^x)^2}$. 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.

- Standardize 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*.*

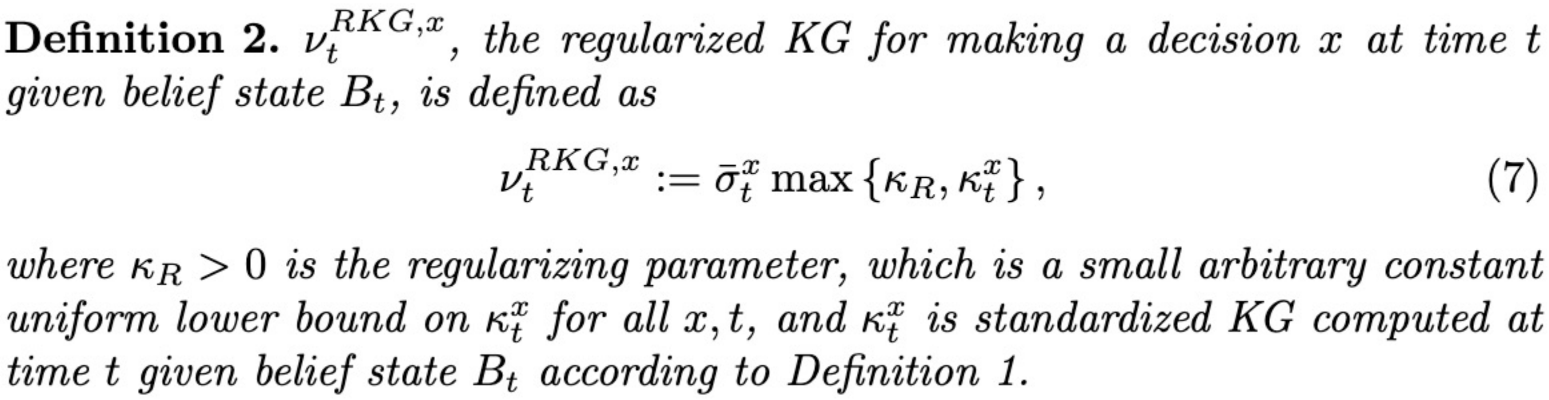
κ_t^x is “standardized” KG, in a sense that it has the same unit as ξ_t^x:

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

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.

Regularization

- Regularize KG



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*

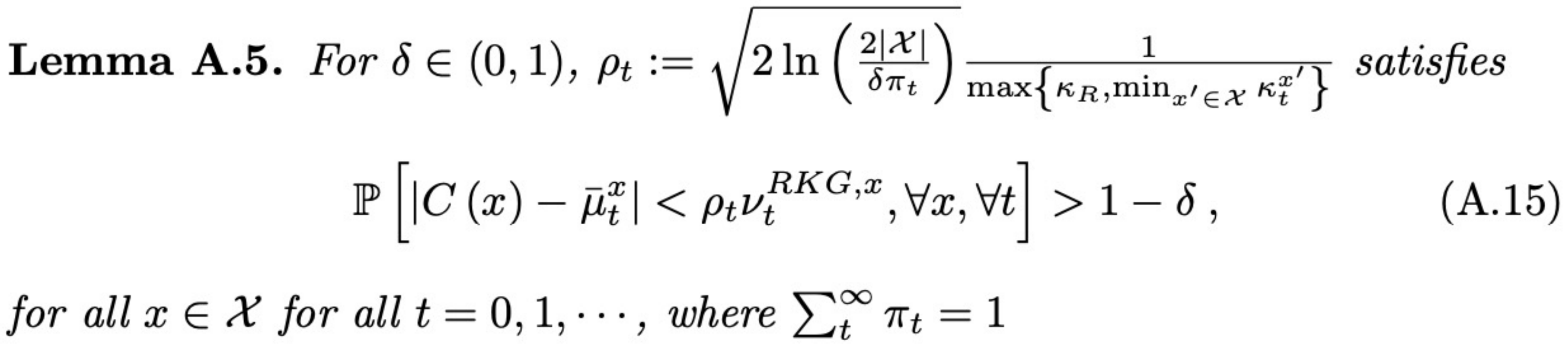
$$\nu_t^{RKG,x} := \bar{\sigma}_t^x \max\left\{\kappa_R, \kappa_t^x\right\}, \tag{7}$$

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.

- Given belief at time t : $C(x) \sim \mathcal{N}(\bar{\mu}_t^x, (\bar{\sigma}_t^x)^2)$
- Where κ_t^x is standardized KG

Bound Step 1

- Bound one-step reward deviation w/ high probability



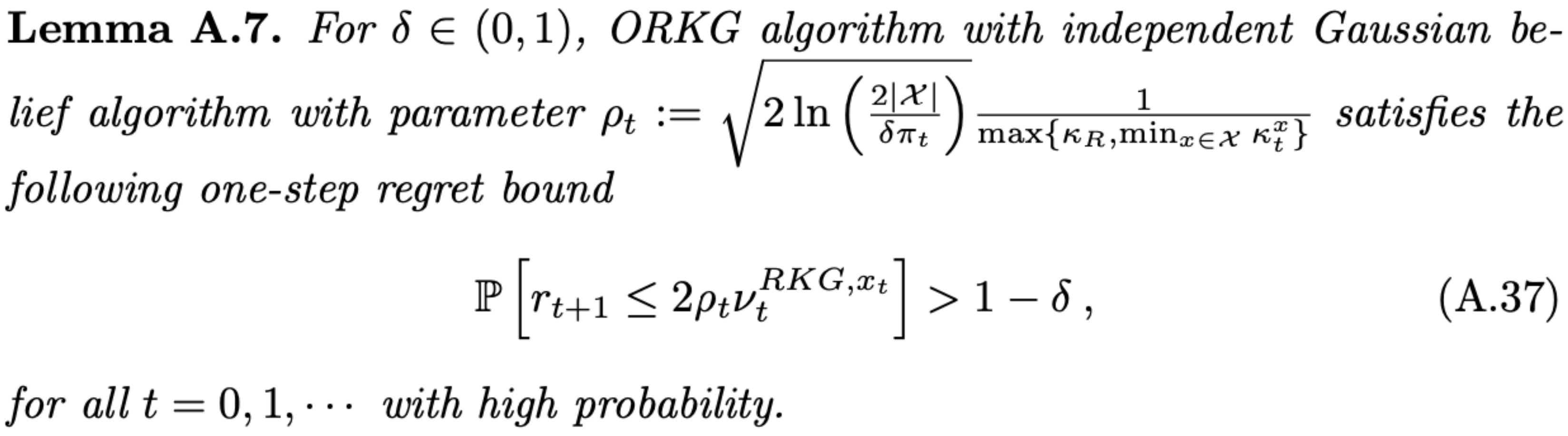
Lemma A.5. *For* $\delta \in (0,1)$, $\rho_t := \sqrt{2\ln\left(\frac{2|\mathcal{X}|}{\delta\pi_t}\right)} \frac{1}{\max\{\kappa_R, \min_{x' \in \mathcal{X}} \kappa_t^{x'}\}}$ *satisfies*

$$\mathbb{P}\left[\left|C(x) - \bar{\mu}_t^x\right| < \rho_t \nu_t^{RKG,x}, \forall x, \forall t\right] > 1 - \delta\,, \qquad \text{(A.15)}$$

for all $x \in \mathcal{X}$ *for all* $t = 0, 1, \cdots$, *where* $\sum_t^{\infty} \pi_t = 1$

Bound Step 2

- Bound one-step regret



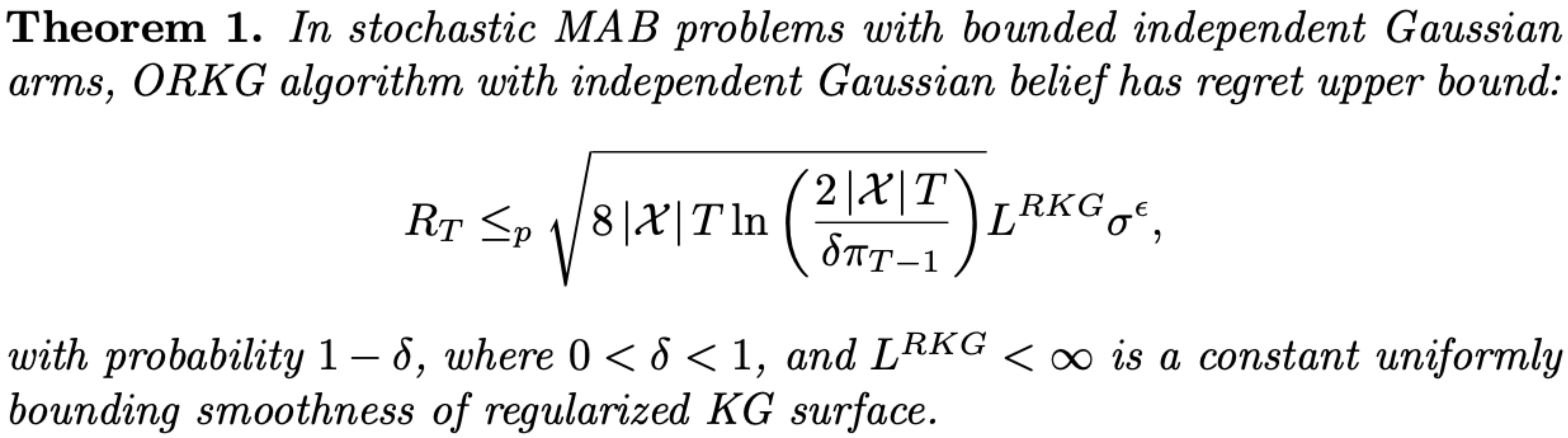
Lemma A.7. *For $\delta \in (0,1)$, ORKG algorithm with independent Gaussian belief algorithm with parameter* $\rho_t := \sqrt{2 \ln\left(\frac{2|\mathcal{X}|}{\delta \pi_t}\right) \frac{1}{\max\{\kappa_R, \min_{x \in \mathcal{X}} \kappa_t^x\}}}$ *satisfies the following one-step regret bound*

$$\mathbb{P}\left[r_{t+1} \le 2\rho_t \nu_t^{RKG,x_t}\right] > 1 - \delta\,, \tag{A.37}$$

for all $t = 0, 1, \cdots$ with high probability.

Bound Step 3

- Bound sum of regrets over T-steps



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.*

Comparison Against Other KG's

- Hyperparams

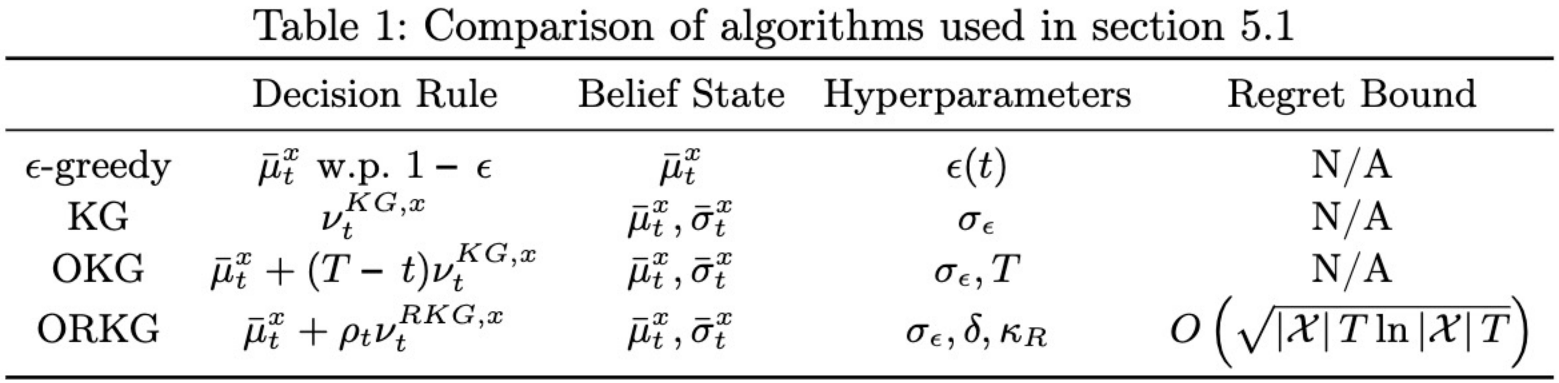


Table 1: Comparison of algorithms used in section 5.1

	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	$\bar{\mu}_t^x + \rho_t \nu_t^{RKG,x}$	$\bar{\mu}_t^x, \bar{\sigma}_t^x$	$\sigma_\epsilon, \delta, \kappa_R$	$O\left(\sqrt{\lvert\mathcal{X}\rvert T \ln \lvert\mathcal{X}\rvert T}\right)$

- Regrets

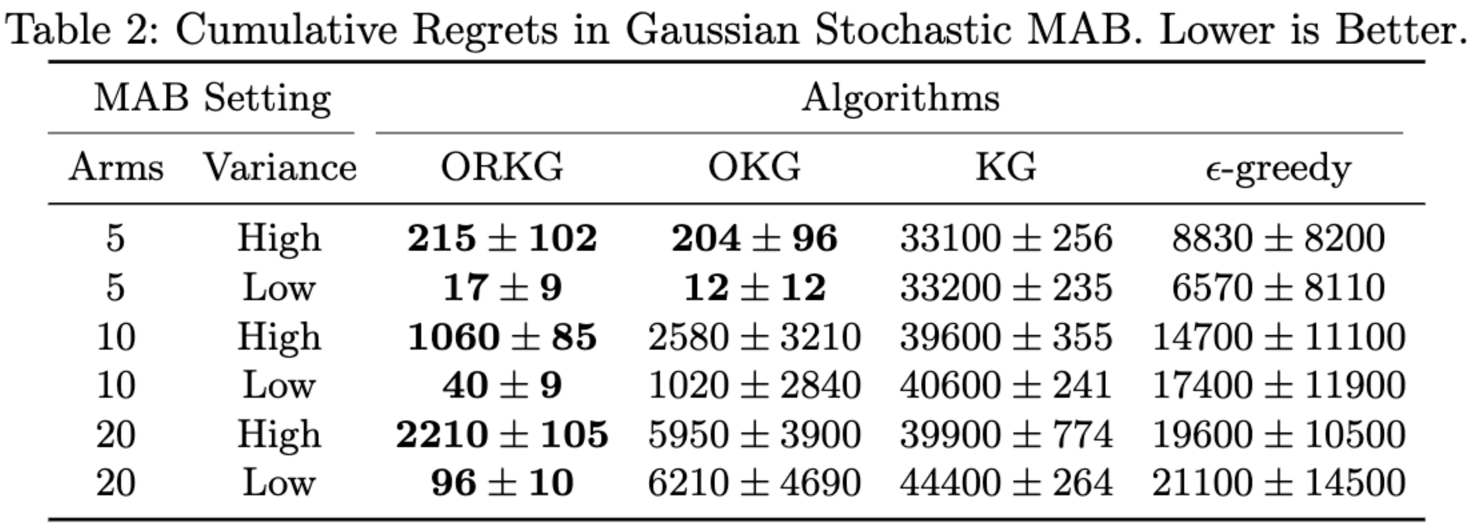


Table 2: Cumulative Regrets in Gaussian Stochastic MAB. Lower is Better.

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

Sensitivity Analysis on κ_R

- Reg. hyperparam.

- Gaussian MAB
  - 10 arms
  - Low variance

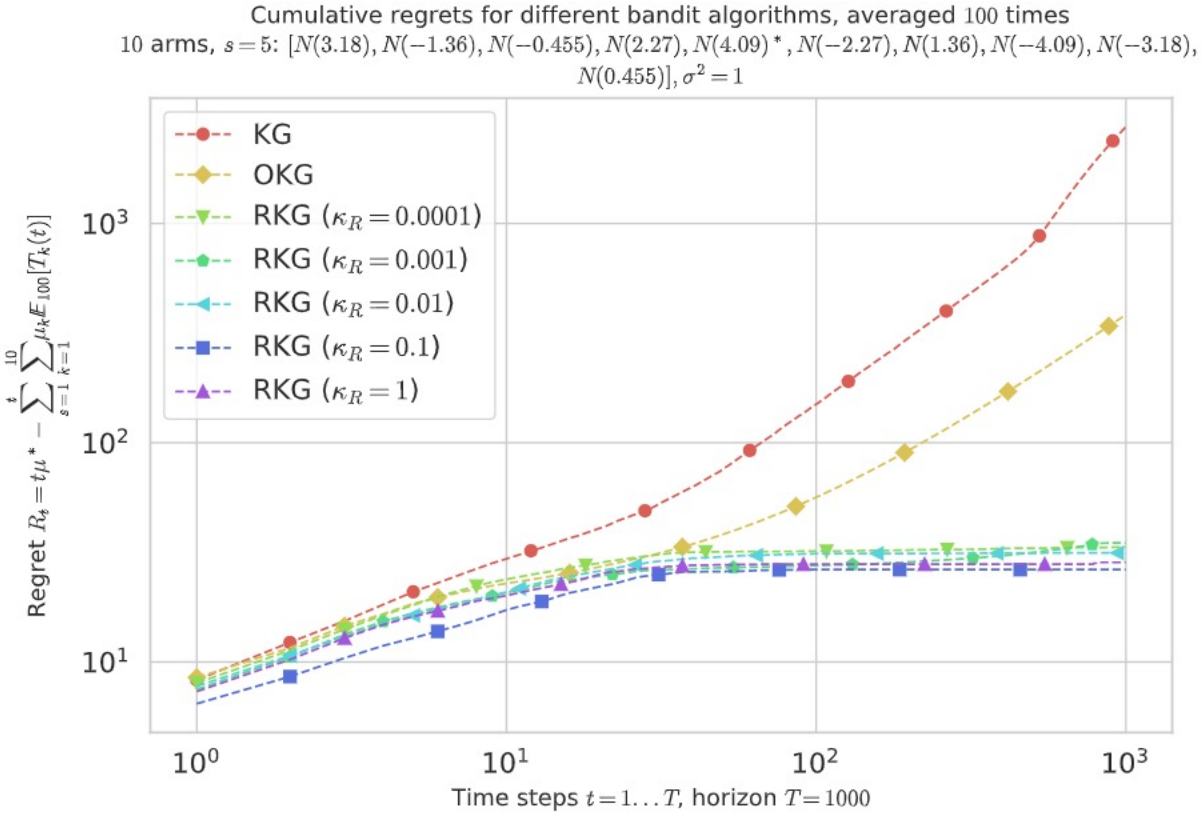


Fig. 1: Sensitivity of ORKG to κ_R in Gaussian MAB ($K = 10, \sigma^2 = 1, \delta = 0.01$).

Sensitivity Analysis on δ

- Probabilistic regret bound hyperparam.
- Gaussian MAB
  - 10 arms
  - Low variance

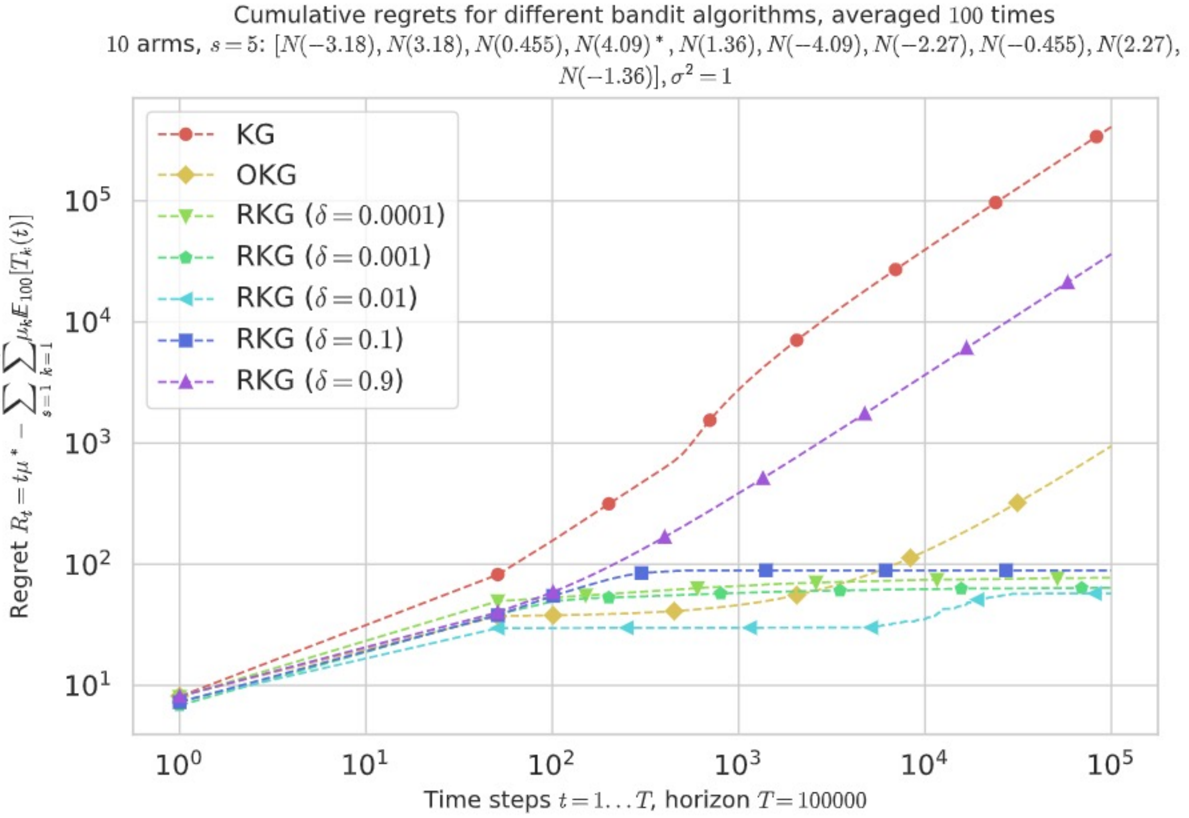
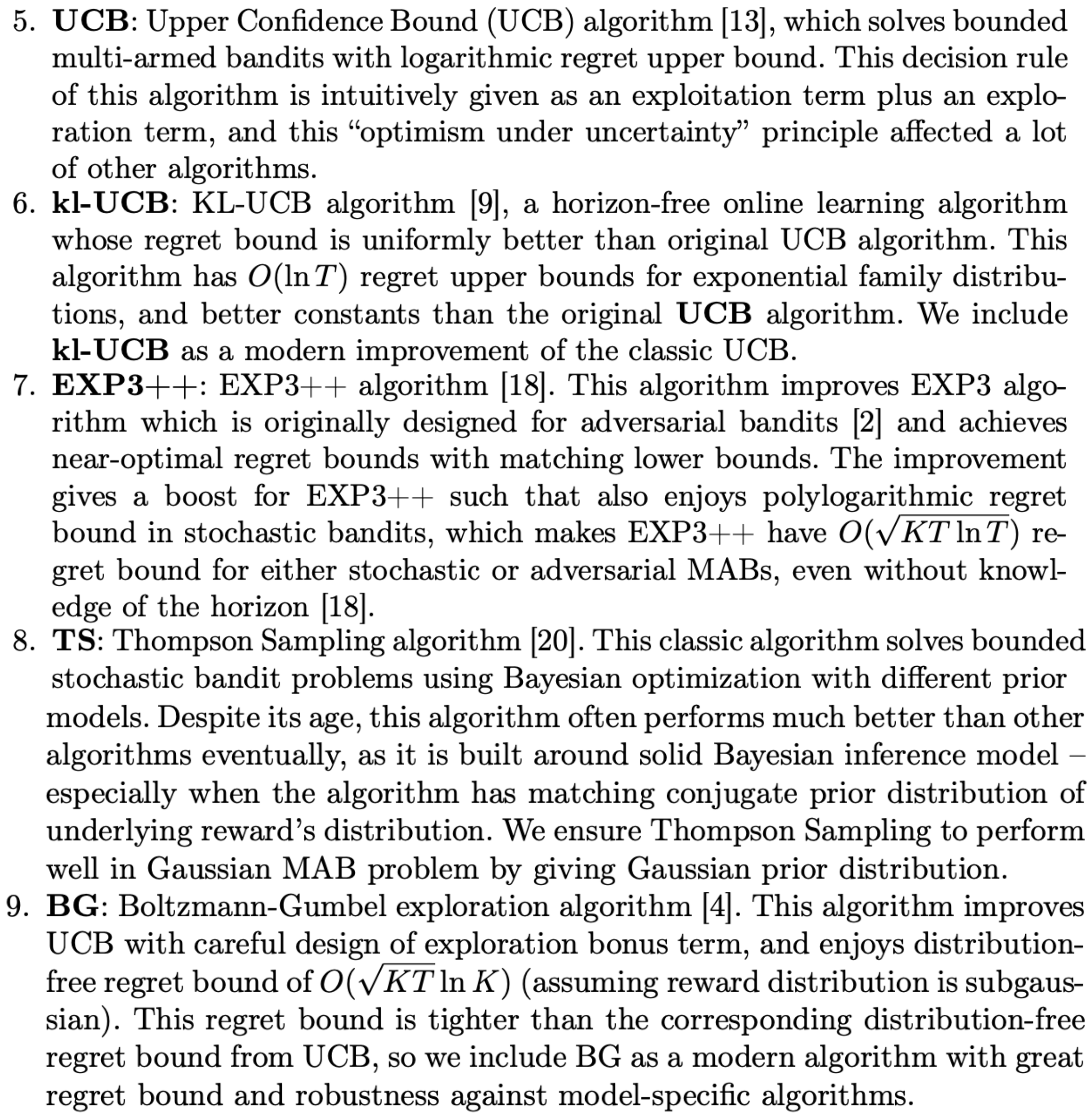


Fig. 2: Sensitivity of ORKG to δ in Gaussian MAB ($K = 10, \sigma^2 = 1, \kappa_R = 0.01$).

MAB Comparison

- Some classics
  - UCB
  - TS
- Some recents
  - kl-UCB
  - EXP3++
  - BG



5. **UCB**: Upper Confidence Bound (UCB) algorithm [13], which solves bounded multi-armed bandits with logarithmic regret upper bound. This decision rule of this algorithm is intuitively given as an exploitation term plus an exploration term, and this "optimism under uncertainty" principle affected a lot of other algorithms.
6. **kl-UCB**: KL-UCB algorithm [9], a horizon-free online learning algorithm whose regret bound is uniformly better than original UCB algorithm. This algorithm has $O(\ln T)$ regret upper bounds for exponential family distributions, and better constants than the original **UCB** algorithm. We include **kl-UCB** as a modern improvement of the classic UCB.
7. **EXP3++**: EXP3++ algorithm [18]. This algorithm improves EXP3 algorithm which is originally designed for adversarial bandits [2] and achieves near-optimal regret bounds with matching lower bounds. The improvement gives a boost for EXP3++ such that also enjoys polylogarithmic regret bound in stochastic bandits, which makes EXP3++ have $O(\sqrt{KT}\ln T)$ regret bound for either stochastic or adversarial MABs, even without knowledge of the horizon [18].
8. **TS**: Thompson Sampling algorithm [20]. This classic algorithm solves bounded stochastic bandit problems using Bayesian optimization with different prior models. Despite its age, this algorithm often performs much better than other algorithms eventually, as it is built around solid Bayesian inference model – especially when the algorithm has matching conjugate prior distribution of underlying reward's distribution. We ensure Thompson Sampling to perform well in Gaussian MAB problem by giving Gaussian prior distribution.
9. **BG**: Boltzmann-Gumbel exploration algorithm [4]. This algorithm improves UCB with careful design of exploration bonus term, and enjoys distribution-free regret bound of $O(\sqrt{KT}\ln K)$ (assuming reward distribution is subgaussian). This regret bound is tighter than the corresponding distribution-free regret bound from UCB, so we include BG as a modern algorithm with great regret bound and robustness against model-specific algorithms.

MAB Comparison

- Regrets

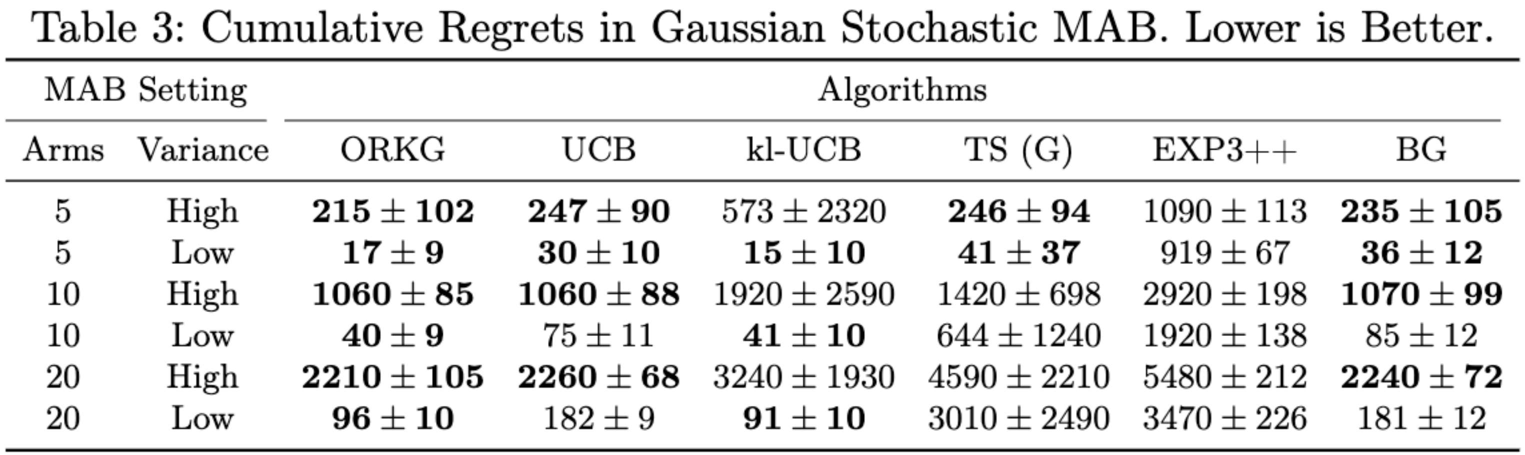
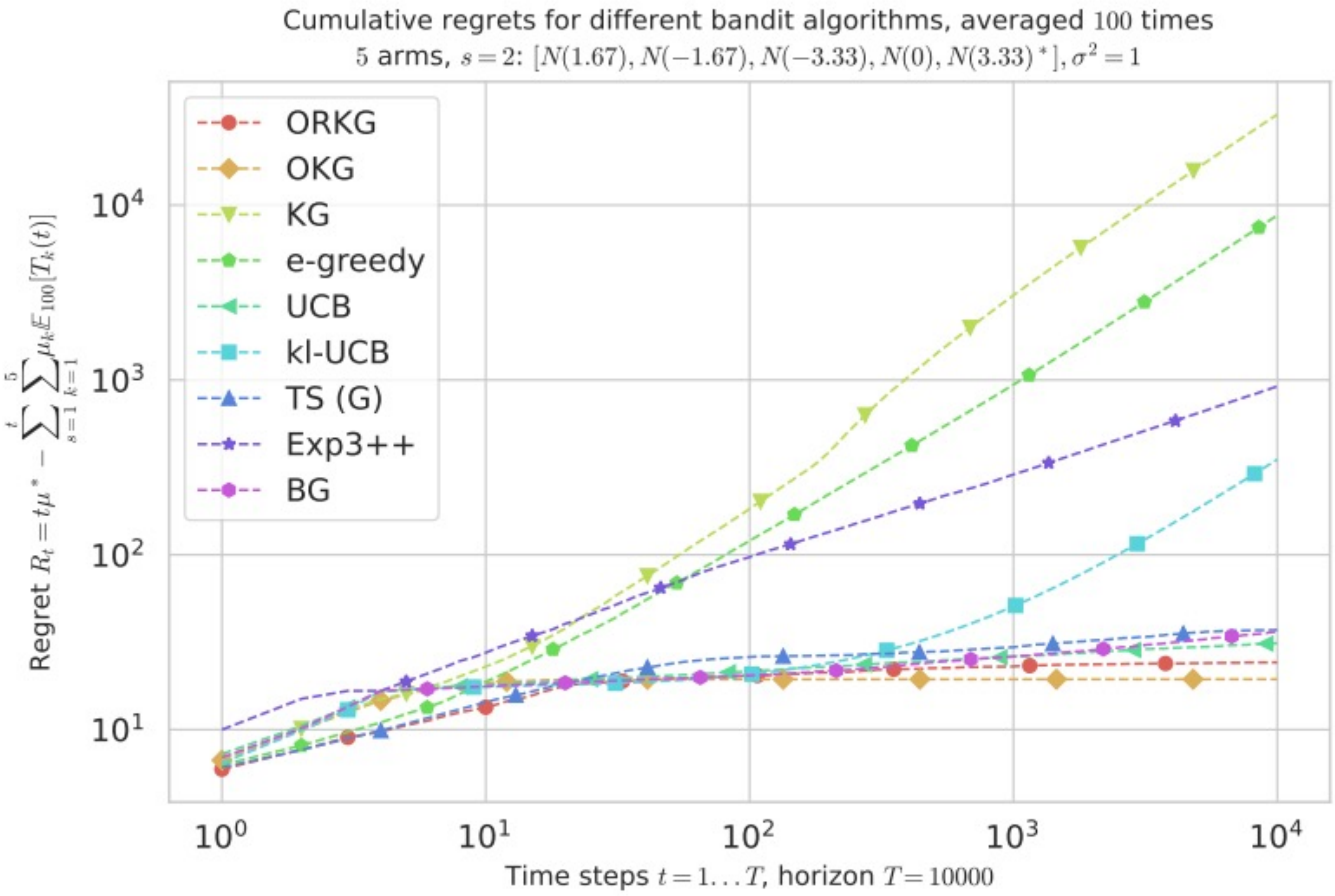


Table 3: Cumulative Regrets in Gaussian Stochastic MAB. Lower is Better.

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

The Regrets

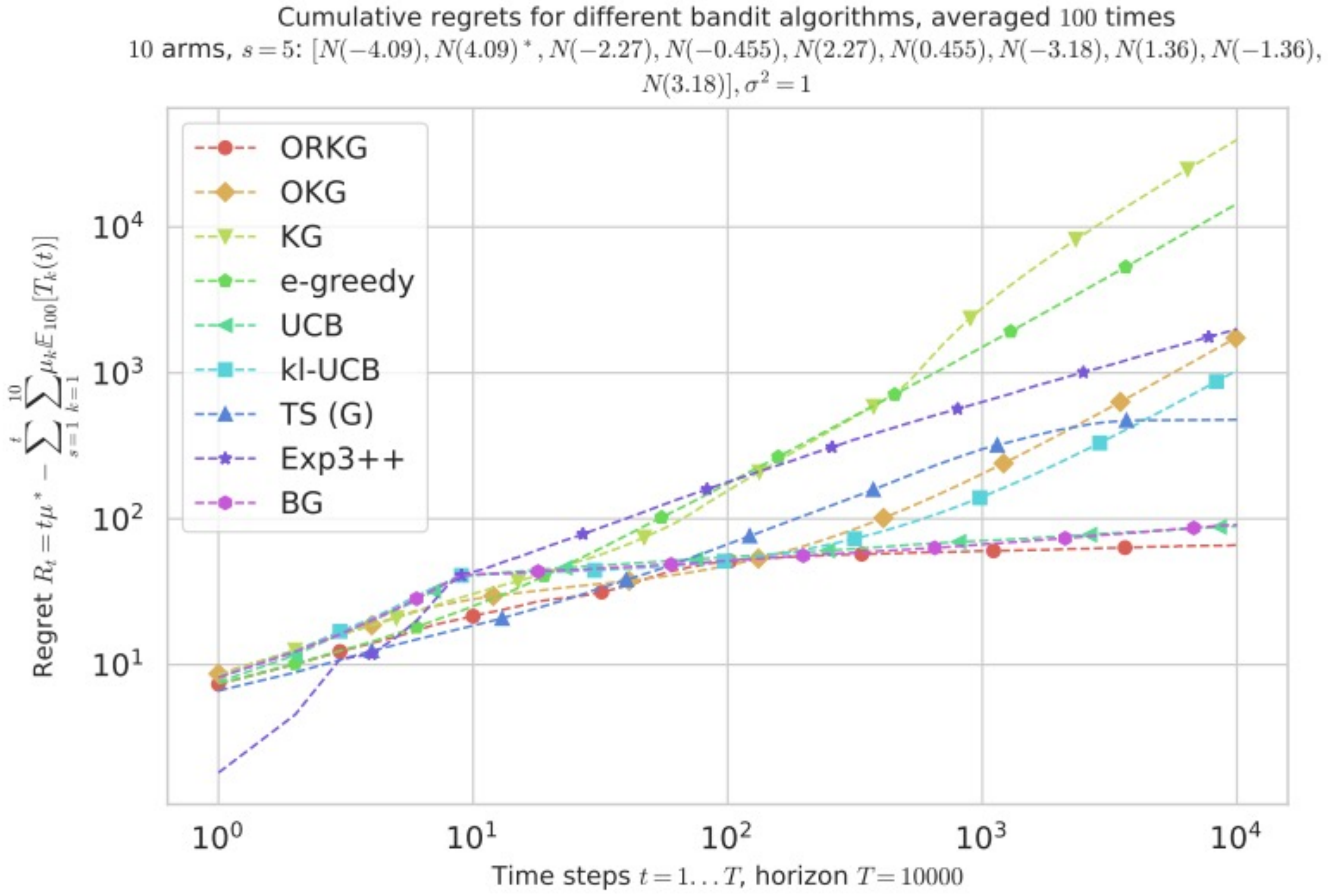
- Gaussian MAB
    - 5 arms
    - Low variance



$$\text{(a) } K = 5, \sigma^2 = 1$$

The Regrets

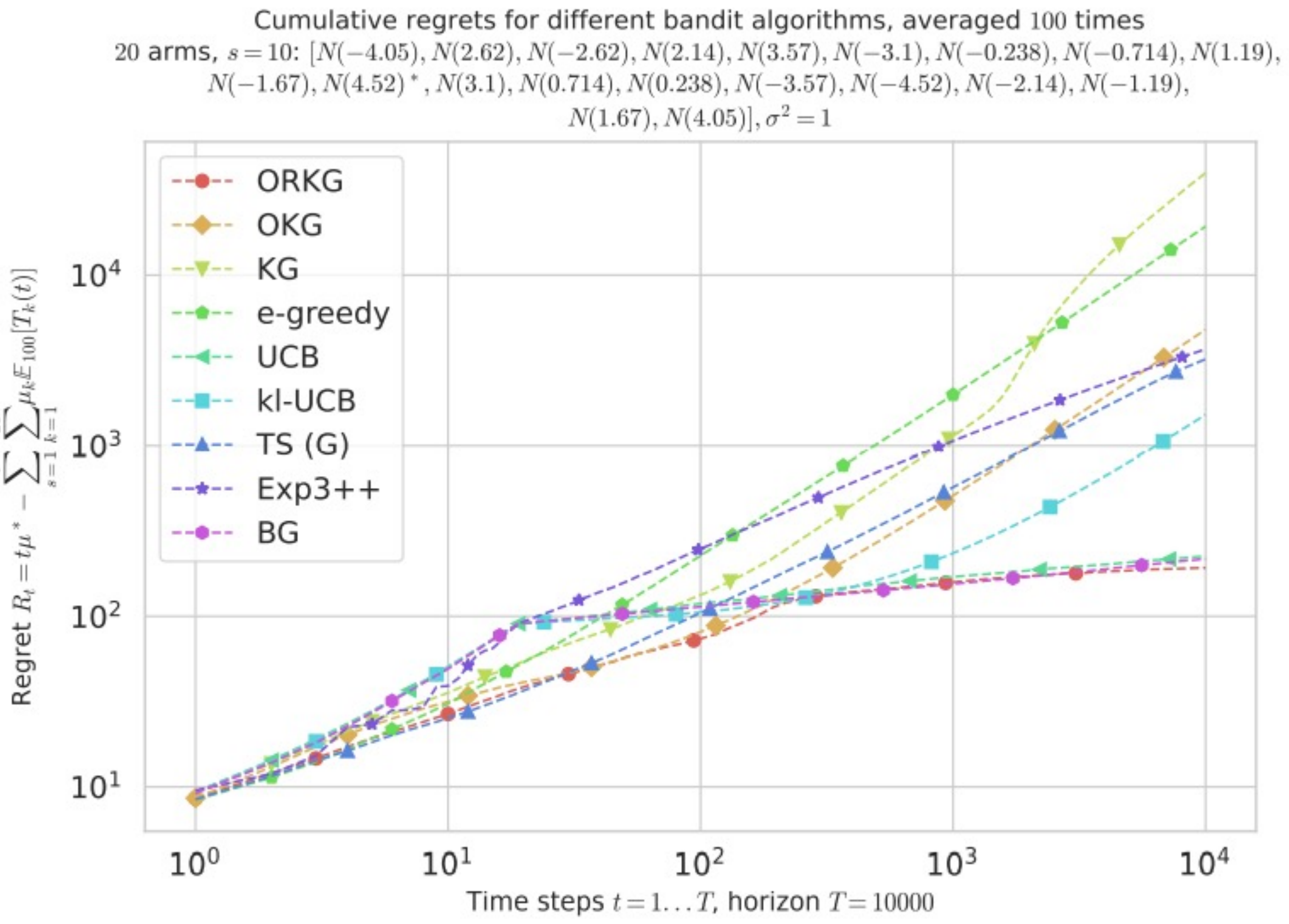
- Gaussian MAB
  - 10 arms
  - Low variance



(c) $K = 10, \sigma^2 = 1$

The Regrets

- Gaussian MAB
  - 20 arms
  - Low variance



$$\text{(e) } K = 20, \sigma^2 = 1$$

Conclusion

- Regularized KG allows regret analysis on KG-based algs
  - First result for KG-based algorithms

- Online Regularized KG algorithm (ORKG), adaptation of regularized KG for MAB problems
  - Has provable sublinear regret bound
  - Shows good performance in Gaussian MAB

Questions?

>> holy@korea.ac.kr