Asymptotically Optimal Policies for Non-Parametric MAB Models Under Generalized Ranking

Michael N. Katehakis

Rutgers University

January 11, 2016

Joint work with Wes Cowan, Rutgers University

Overview

- Non-Parametric MAB Framework
- What Makes a Policy Good?
- 'Idealized Assumptions'
- How Good is Great?
- Policy π^* (UCB-$(\mathcal{F}, s, \tilde{d})$)
- Applications:
    - Separable Pareto Models
    - General Uniform Models
    - Three Normal Examples
- References

A General Framework for MABs

- Known family of densities $\mathcal{F}$
- Controller faces N unknown 'bandits':

$$\underline{f} = \{f_1, f_2, \dots, f_N\} \subset \mathcal{F}$$

- May sample $i.i.d.$ from any bandit: $X_1^i, X_2^i, \dots \sim f_i$
- Given t samples of i, may construct estimator $\hat{f}_t^i$
- Sequential sampling policy π, $\pi(n) = i$ samples i at time n
    - $T_\pi^i(n)$: # of samples of i at global time n (global n vs local $T_\pi^i(n)$)
- Score functional $s : \mathcal{F} \mapsto \mathbb{R}$.
- Optimal bandits: $s(f_{i^*}) = s^* = \max_j s(f_j)$.

General Goal:

- A policy π that samples optimal bandits as often as possible
- Efficiently balance exploration vs exploitation

What is Good?

Let $\mathcal{O}(\underline{f}) = \{i : s(f_i) = s^*(\underline{f})\}, \quad B(\underline{f}) = \{i : s(f_i) < s^*(\underline{f})\}$

be the set of optimal, sub-optimal bandits

Basic Principle: *Activations of optimal bandits cannot be regretted.*

Definition (Uniformly Fast Policies.)

A policy π is *Uniformly Fast* if,for all $\underline{f} = (f_i)$, $f_i \in \mathcal{F}$, $\alpha > 0$

$$\sum_{i \in B(\underline{f})} \mathbb{E}_{\underline{f}} \left[T_\pi^i(n) \right] = o(n^\alpha),$$

- Regret:

$$R_\pi(n) = R_\pi(n; \underline{f}) \sum_{i \in B(\underline{f})} (s^*(\underline{f}) - s(f_i)) \mathbb{E}_{\underline{f}} \left[T_\pi^i(n) \right]$$

- Robbins (1952), Lai and Robbins (1985), Katehakis and Robbins (1995), Burnetas and Katehakis (1996), Honda and Takemura (2010), Honda and Takemura (2011), Honda and Takemura (2013)

Structure of Bandit Space

- KL-Divergence as 'distance/similarity' in $\mathcal{F}$:

$$\mathbf{I}(f, g) = \mathbb{E}_f \left[\ln \left(\frac{f(X)}{g(X)} \right) \right].$$

- $\mathbf{I}(f, g) = 0$ implies $f = g$ (a.e.)
- $\mathbf{I}(f, g) < \infty$ implies g supports f (w.p. 1)
- Note: not a true metric - that's okay!
- $\mathcal{F}$ characterized by

$$\mathbb{K}_f(\rho) = \inf_{g \in \mathcal{F}} \{ \mathbf{I}(f, g) : s(g) > \rho \}.$$

  - $\mathbb{K}_f(\rho)$: Distance to nearest ρ-better g

So Good, No Better

Assume the following conditions hold, for any $f \in \mathcal{F}$, and all $\epsilon, \delta > 0$.

- **Condition B1:** $\forall\, f \in \mathcal{F}, \rho \in s(\mathcal{F})$, $\exists \tilde{f} \in \mathcal{F} : s(\tilde{f}) > \rho$ and $\mathbf{I}(f, \tilde{f}) < \infty$.
- **Condition B2:** s is continuous at each $f \in \mathcal{F}$, with respect to $\mathbf{I}$.

Theorem (Lower Bound on Sub-Optimal Activations.)

For any $(\mathcal{F}, s)$ that satisfy: B1 & B2.
Then, $\forall \pi\ UF$ and all $\underline{f}$, the following holds for each sub-optimal i:

$$\liminf_{n} \frac{\mathbb{E}_{\underline{f}}\left[T_{\pi}^{i}(n)\right]}{\ln n} \geq \frac{1}{\mathbb{K}_{f_i}(s^*)} \ .$$

Are there policies ('asymptotically optimal') that achieve this lower bound?

Realizing the Bound

Goal: construct policies π, based on knowledge of $\mathcal{F}$ and s, that achieve this lower bound, that is for all sub-optimal i:

$$\lim_n \mathbb{E}[T^i_\pi(n)]/\ln n = 1/\mathbb{K}_{f_i}(s^*)$$

Let ν be a (context-specific) measure of similarity of $\mathcal{F}$.
Assume the following conditions hold, for any $f \in \mathcal{F}$, and all $\epsilon, \delta > 0$.

- **Condition R1:** $\mathbb{K}_f(\rho)$ is continuous w.r.t ρ, and w.r.t f under ν.
- **Condition R2:** $\mathbb{P}_f(\nu(\hat{f}_t, f) > \delta) \leq o(1/t)$.
- **Condition R3:** For some sequence $d_t = o(t)$ (independent of ϵ, δ, f),

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f) - \epsilon)) \leq e^{-\Omega(t)} e^{-(t-d_t)\delta},$$

where the dependence on ϵ and f are suppressed into the $\Omega(t)$ term.

Standard notation: $o(n)$, $O(n)$ and $\Omega(n)$ denote a function $h(n)$ with the following properties respectively. i) $\lim_n h(n)/n = 0$. ii) $\exists\ c > 0$ and $n_0 \geq 1$ such that $h(n) \leq c\,n$, for all $n > n_0$. iii) $\exists\ c > 0$ and $n_0 \geq 1$ such that $h(n) \geq c\,n$, for all $n > n_0$.

Discussion

◇ **Condition R1:** $\mathbb{K}_f(\rho)$ is continuous w.r.t ρ, and w.r.t f under ν.
It characterizes, in some sense, the structure of $\mathcal{F}$ as smooth.
To the extent that $\mathbb{K}_f(\rho)$ *can* be thought of as a Hausdorff distance on $\mathcal{F}$, Condition R1 restricts the "shape" of $\mathcal{F}$ relative to s.

◇ **Condition R2:** $\mathbb{P}_f(\nu(\hat{f}_t, f) > \delta) \leq o(1/t)$.
The estimators $\hat{f}_t$ are "honest" and converge to f sufficiently quickly with t.

◇ **Condition R3:** For some sequence $d_t = o(t)$ (independent of ϵ, δ, f),

$$\mathbb{P}_f(\delta < \mathbb{K}_{\hat{f}_t}(s(f) - \epsilon)) \leq e^{-\Omega(t)} e^{-(t-d_t)\delta},$$

It often seems to be satisfied by $\hat{f}_t$ converging to f sufficiently quickly, as well as $\hat{f}_t$ being "useful", in that $s(\hat{f}_t)$ converges sufficiently quickly to $s(f)$.
The form of the above bound, while specific in its dependence on t and δ, can be relaxed somewhat, but such a bound frequently seems to exist in practice, for natural choices of $\hat{f}_t$.

Policy UCB-$(\mathcal{F}, s, \hat{f}_t, \tilde{d})$

- Let $\hat{f}_t^i$ be an estimator of f_i given t i.i.d. samples.
- Let $\tilde{d}(t) > 0$ be a non-decreasing function with $\tilde{d}(t) = o(t)$.
- Define, for any t such that $t > \tilde{d}(t)$, the following index function:

$$u_i(n,t) = \sup_{g \in \mathcal{F}} \left\{ s(g) : \mathbf{I}(\hat{f}_t^i, g) \leq \frac{\ln n}{t - \tilde{d}} \right\},$$

Policy π^* (UCB-$(\mathcal{F}, s, \tilde{d})$):

i) For $n = 1, 2, \ldots, n_0 \times N$, sample each bandit n_0 times, and

ii) for $n \geq n_0 \times N$, sample from bandit

$$\pi^*(n+1) = \arg\max_i u_i\left(n, T_{\pi^*}^i(n)\right),$$

breaking ties uniformly at random

Intuition: Activate according to best score within plausible distance of best bandit estimate.

Related: (Burnetas and Katehakis 1996): (Auer and Ortner 2010), (Cappé, Garivier, Maillard, Munos, and Stoltz 2013)

Finite Horizon Bound for Activations under Policy π^*

Theorem

For any sub-optimal i *and any optimal* i^**, and*

◇ $\forall \epsilon > 0$ *such that* $s^* - \epsilon > s(f_i)$*,*

◇ $\forall \delta > 0$ *such that* $\inf_{g \in \mathcal{F}} \{\mathbb{K}_g(s^* - \epsilon) : \nu(g, f_i) \leq \delta\} > 0$*:*

$$\mathbb{E}\left[T^i_{\pi^*}(n)\right] \leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{\mathbb{K}_g(s^* - \epsilon) : \nu(g, f_i) \leq \delta\}} + o(\ln n)$$
$$+ \sum_{t=n_0 N}^{n} \mathbb{P}\left(\nu(\hat{f}^i_t, f_i) > \delta\right)$$
$$+ \sum_{t=n_0 N}^{n} \sum_{k=n_0}^{t} \mathbb{P}\left(u_{i^*}(t, k) \leq s^* - \epsilon\right).$$

Asymptotic Optimality

Theorems 1 and 2 lead to the following theorem:

Theorem

Let $(\mathcal{F}, s, \hat{f}_t, \nu)$ satisfy Conditions B1, B2 & R1 - R3.
Let $d = \{d_t\}$ be as in Condition R3 and
$\tilde{d}(t) - d_t \geq \Delta > 0$ for some Δ, for all t, then

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\mathbb{K}_{f_i}(s^*)},\ \forall\, \underline{f} \in \mathcal{F},\ and\ \forall i\ suboptimal.$$

Applications: Separable Pareto Models

$$\mathcal{F}_\ell = \left\{ f_{\alpha,\beta}(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}} \text{ for } x \geq \beta : \ell < \alpha < \infty, \beta > 0 \right\}$$

$X \sim$ Pareto(α, β), X is distributed over $[\beta, \infty)$,
with $\mathbb{E}[X] = \alpha\beta/(\alpha - 1)$ if $\alpha > 1$, and $\mathbb{E}[X]$ as infinite or undefined if $\alpha \leq 1$.
We are interested in $\mathcal{F}_0$, the family of unrestricted Pareto distributions, and $\mathcal{F}_1$, the family of Pareto distributions with finite means.

A score function $s(\alpha, \beta) = s(f_{\alpha,\beta})$ of interest should be an increasing function of β, and a decreasing function of α.

We consider score functions:

$$s(f) = s(\alpha, \beta) = a(\alpha)b(\beta)$$

where we take a to be a positive, continuous, decreasing, invertible function of α for $\alpha > \ell$, and b to be a positive, continuous, non-decreasing function of β^*.

* When the goal is to obtain large rewards from the bandits activated, there are two effects of interest: rewards from a given bandit will be biased towards larger values for decreasing α and increasing β.

Applications: Separable Pareto Models - Continued

This general Pareto model of $s(\alpha, \beta) = a(\alpha)b(\beta)$, includes several natural score functions of interest, in particular:

i) In the case of the restricted Pareto distributions with finite mean, we may take s as the expected value, and

$$s(\alpha, \beta) = \alpha\beta/(\alpha - 1),$$

with $a(\alpha) = \alpha/(\alpha - 1)$ and $b(\beta) = \beta$.

ii) For unrestricted Pareto distributions, the score function

$$s(\alpha, \beta) = 1/\alpha,$$

leads to the controller's goal to be to find the bandit with minimal α. In this case, $a(\alpha) = 1/\alpha$ and $b(\beta) = 1$. Can be used in comparing the asymptotic tail distributions of bandits, $\mathbb{P}(X \geq k)$ as $k \to \infty$, or the conditional restricted expected values, $\mathbb{E}[X|X \leq k]$ as $k \to \infty$.

iii) A third score function

$$s(\alpha, \beta) = \beta 2^{1/\alpha},$$

with $a(\alpha) = 2^{1/\alpha},\ b(\beta) = \beta$, can be used for the median, defined over unrestricted Pareto distributions.

Applications: Separable Pareto Models - Continued

◇ Assume: $a(\alpha) \to \infty$ as $\alpha \to \ell$.
This guarantees that Condition B1 is satisfied by s.

◇ For $f = f_{\alpha,\beta} \in \mathcal{F}_\ell$, and a sample of size t of i.i.d. samples under f, take the estimator $\hat{f}_t = f_{\hat{\alpha}_t,\hat{\beta}_t}$ where

$$\begin{aligned} \hat{\beta}_t &= \min_{n=1,\ldots,t} X_k, \\ \hat{\alpha}_t &= \frac{t-1}{\sum_{k=1}^{t} \ln\left(\frac{X_k}{\hat{\beta}_t}\right)}. \end{aligned} \tag{1}$$

Define the following functions, $L^+(\delta)$, $L^-(\delta)$, as the smallest and largest positive solutions to $L - \ln L - 1 = \delta$ for $\delta \geq 0$, respectively.
$L^-(\delta)$ may be expressed in terms of the Lambert-W function, $L^-(\delta) = -W(e^{-1-\delta})$, taking $W(x)$ be the principal solution to $We^W = x$ for $x \in [-1/e, \infty)$. An important property will be that $L^\pm(\delta)$ is continuous as a function of δ, and $L^\pm(\delta) \to 1$ as $\delta \to 0$.

Policy $\pi^*_{\mathsf{P,s}}$ (UCB-PARETO)

i) For $n = 1, 2, \dots 3N$, sample each bandit 3 times, and

ii) for $n \geq 3N$, sample from bandit $\pi^*_{\mathsf{P,s}}(n+1) = \arg\max_i u_i \left(n, T^i_{\pi^*_{\mathsf{P,s}}}(n)\right)$ breaking ties uniformly at random, where

$$u_i(n,t) = \begin{cases} \infty & \text{if } \hat{\alpha}^i_t L^- \left(\frac{\ln n}{t-2}\right) \leq \ell, \\ b\left(\hat{\beta}^i_t\right) a \left(\hat{\alpha}^i_t L^- \left(\frac{\ln n}{t-2}\right)\right) & \text{else.} \end{cases}$$

Theorem

Policy $\pi^*_{P,s}$ as defined above is asymptotically optimal: for each sub-optimal bandit i the following holds:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*_{P,s}}(n)\right]}{\ln n} = \frac{1}{\frac{1}{\alpha_i} a^{-1}\left(\frac{s^*}{b(\beta_i)}\right) - \ln\left(\frac{1}{\alpha_i} a^{-1}\left(\frac{s^*}{b(\beta_i)}\right)\right) - 1}.$$

Applications: General Uniform Models

$$\mathcal{F} = \left\{ f_{a,b}(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b : -\infty < a < b < \infty \right\}$$

- General Model of Interest: $s(f) = s(a, b)$.
    - $s(a, b)$: continuous, increasing function of a
    - $s(a, b)$: continuous, increasing function of b

Contains standard case of interest:

$$s_\mu(a, b) = (a + b)/2.$$

Applications: General Uniform Models - Continued

$$\mathcal{F} = \left\{ f_{a,b}(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b : -\infty < a < b < \infty \right\}$$

- Estimators of $f = f_{a,b}$ as $\hat{f}_t = f_{\hat{a}_t, \hat{b}_t}$ where

$$\hat{a}_t = \min_{n=1,\dots,t} X_n \quad \hat{b}_t = \max_{n=1,\dots,t} X_n.$$

- Index of Optimal Policy π^*, $\tilde{d} = 2$:

$$u_i(n,t) = s(\hat{a}_t^i, \hat{a}_t^i + n^{\frac{1}{t-2}}(\hat{b}_t^i - \hat{a}_t^i)).$$

with the particular case for s_μ:

$$u_i(n,t) = \hat{a}_t^i + \frac{1}{2} n^{\frac{1}{t-2}}(\hat{b}_t^i - \hat{a}_t^i).$$

Applications: General Uniform Models - Continued

Policy π^* is asymptotically optimal, and for all $\{f_i = f_{a_i,b_i}\} \subset \mathcal{F}$, for all sub-optimal i:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\min_{b_i \leq b}\{\ln(b - a_i) : s(a_i, b) \geq s^*\} - \ln(b_i - a_i)}.$$

with the particular case for s_μ:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{1}{\ln\left(\frac{2s^* - 2a_i}{b_i - a_i}\right)}.$$

Applications: Normal, Unknown μ_i, σ_i, Maximize Mean

$$\mathcal{F} = \left\{ f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} : -\infty < \mu < \infty, 0 < \sigma < \infty \right\}$$

- Score functional: $s(f_{\mu,\sigma}) = \mathbb{E}_f\left[X\right] = \mu$.
- Standard Estimators: $\hat{\mu}_t$ and $\hat{\sigma}_t^2$.
- Index of Optimal Policy $\pi^* = \pi_{\mathsf{CHK}}$, $\tilde{d} = 2$:

$$u_i(n,t) = \hat{\mu}_t^i + \hat{\sigma}_t^i \sqrt{n^{\frac{2}{t-2}} - 1}.$$

- Asymptotic Optimality: For all $\{f_i = f_{\mu_i,\sigma_i}\} \subset \mathcal{F}$, for sub-optimal i:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi_{\mathsf{CHK}}}(n)\right]}{\ln n} = \frac{2}{\ln\left(1 + \frac{(\mu^* - \mu_i)^2}{\sigma_i^2}\right)}.$$

(Cowan, Honda, and Katehakis 2015)

Normal, Minimize Variance, known μ_i,

$$\mathcal{F}_M = \left\{ f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M)^2}{2\sigma^2}} : 0 < \sigma < \infty \right\}$$

- Score functional: $s(f_\sigma) = 1/\mathsf{Var}_f(X) = 1/\sigma^2$.
- Standard Estimators: $\hat{\sigma}_t^2$.
- Index of Optimal Policy π^*, $\tilde{d} = 2$:

$$u_i(n,t) = L^+ \left(\frac{2 \ln n}{t-2} \right) / (\hat{\sigma}_t^i)^2$$

  with $L^+(\delta)$ largest positive solution: $L - \ln L - 1 = \delta$.
- Asymptotic Optimality: For all $\{f_i = f_{\sigma_i}\} \subset \mathcal{F}_M$, for sub-optimal i:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{2}{\frac{\sigma_i^2}{\sigma_*^2} - \ln\left(\frac{\sigma_i^2}{\sigma_*^2}\right) - 1}.$$

Normal κ-Threshold Probability, known σ_i

$$\mathcal{F}_i = \left\{ f_{\mu,\sigma_i}(x) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma_i^2}} : -\infty < \mu < \infty \right\}$$

- Score functional: $s(f_{\mu,\sigma}) = \mathbb{P}_f(X > \kappa) = 1 - \Phi((\kappa - \mu)/\sigma)$.
- Standard Estimators: $\hat{\mu}_t$.
- Index of Optimal Policy π^*, $\tilde{d} = 1$:

$$u_i(n,t) = 1 - \Phi\left(\frac{\kappa - \hat{\mu}_t^i}{\sigma_i} - \sqrt{\frac{2\ln n}{t-1}} \right).$$

- Asymptotic Optimality: For all $\{f_i = f_{\mu_i,\sigma_i} \in \mathcal{F}_i\}$, for sub-optimal i:

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^*}(n)\right]}{\ln n} = \frac{2}{\left(\frac{\kappa - \mu_i}{\sigma_i} - \Phi^{-1}\left(1 - p^*\right)\right)^2}.$$

Final Comments: Past and Current Work


Lai - Robbins 1985: $f(x;\theta_i)$ unknown 1-dim (scalar) $\theta_i \in \Theta$

$s(f_i) = \mu(\theta_i)$, $\mu(\theta^*) = \max_i\{\mu(\theta_i)\}$

$$\mathbb{K}_i^{\mathsf{LR}}(\theta^*) = \mathbb{I}(\theta_i, \theta^*)$$

Burnetas - Katehakis 1996: $f(x;\underline{\theta}_i)$ unknown multi-dim (vector) $\underline{\theta}_i \in \underline{\Theta}$

$s(f_i) = \mu(\underline{\theta}_i)$, $\mu(\underline{\underline{\theta}}^*) = \max_i\{\mu(\underline{\theta}_i)\}$ $\qquad \underline{\underline{\theta}} = (\underline{\theta}_1, \ldots, \underline{\theta}_N)$

$$\mathbb{K}_i^{\mathsf{BK}}(\underline{\underline{\theta}}^*) = \inf_{\underline{\theta}_i' \in \underline{\Theta}_i} \{\mathbf{I}(\underline{\theta}_i, \underline{\theta}_i') : \mu(\underline{\theta}_i') > \mu(\underline{\theta}^*)\}$$

Cowan - Katehakis 2015: $f_i \in \mathcal{F}_i$

$s^* = \max_j\{\, s(f_j)\,\}$

$$\mathbb{K}_i^{\mathsf{CK}}(s^*) = \inf_{g \in \mathcal{F}_i} \{\mathbf{I}(f_i, g) : s(g) > s^*\}$$

For all the above, under conditions analogous to B1, B2, $\forall\, \underline{f} \in \underline{\mathcal{F}}$, *and* $\forall i$ *suboptimal*

$$\liminf_n \frac{\mathbb{E}\left[T_\pi^i(n)\right]}{\ln n} \geq \frac{1}{\mathbb{K}_i(s^*)},\ \forall\ UF\ \pi$$

Asymptotically Optimal (Efficient) Policies of Lai - Robbins 1985

LR-UM Policies ϕ^* conditions 3.1-3.3 of L-R (1985)
At time n define: $\pi^{LR}(n)$

- Take $\{a_{ni}\}$ positive sequences of constants that satisfy regularity conditions in L-R(1985)
- at $n = 1, \ldots, N$ sample form Π_n (initial sampling)
- sample mean estimates: $\hat{\mu}_i(n) = \mu(\hat{\theta}_i)$
- First UCBs: $g_n^i(\hat{\theta}_i)(= u_i^{LR}(\hat{\theta}_i)) = \inf_\lambda\{\lambda > \mu(\hat{\theta}_i) \ : \ \mathbb{I}(\hat{\theta}_i, \lambda) \geq a_{ni}\}$
- Take a $\delta \in (0, 1/N)$
- for $n+1 > N$ compute: j: $n+1 = mN + j$ and j_n^*:

$$\hat{\mu}_{j_n^*} = \max\{\hat{\mu}_i(n) \ : \ T^i_{\pi^{LR}}(n) > \delta n\}$$

  and

- $\pi^{LR}(n+1) = \begin{cases} j & \text{if } \hat{\mu}_{j_n^*} < g_n^j(\hat{\theta}_i)) \\ j_n^* & \text{otherwise} \end{cases}$
- Then

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^{LR}}(n)\right]}{\ln n} = \frac{1}{\mathbb{K}_i(\mu^*)}, \ \forall \ non-optimal \ i$$

Asymptotically Optimal (Efficient) Policies of Burnetas Katehakis 1996

BK-UM Policies π^* under conditions A1, A2, A3 of B-K (1996)
At time n define: $\pi^{BK}(n)$

- Take some initial samples from each population, so that at round n so that $T^i_{\pi^*}(n) > 0$ for all i. Initial estimates $\hat{\underline{\theta}}_i(n) = \hat{\underline{\theta}}_i(T^i_{\pi^*}(n))$
- 2-nd UCBs:

$$u_i^{BK}(n,t) = sup_{\underline{\theta}'_i \in \underline{\Theta}_i} \left\{ \mu(\underline{\theta}'_i) \ : \ \mathbb{I}(\hat{\underline{\theta}}_i, \underline{\theta}'_i) < \frac{\ln n}{t} \right\}$$

- 2-nd UCB index based Efficient Policies:

$$\pi^{BK}(n+1) = \arg\max_i \left\{ u_i^{BK}\left(n, T^i_{\pi^*}(n)\right) \right\}$$

breaking ties uniformly at random

- Then

$$\lim_n \frac{\mathbb{E}\left[T^i_{\pi^{BK}}(n)\right]}{\ln n} = \frac{1}{\mathbb{K}_i(\mu(\theta^*))}, \ \forall \ non-optimal \ i$$

π^* is a pure index policy

Asymptotically Optimal Policies of Cowan and Katehakis 2015

Policy UCB-$(\mathcal{F}, s, \hat{f}_t, \tilde{d}, \nu)$ π^* under conditions B1-B2 & R1-R3 of C+K (2015)

- Let $\tilde{d}(t) > 0$ be a non-decreasing function with $\tilde{d}(t) = o(t)$
- For $n = 1, 2, \ldots, n_0 \times N$, sample each bandit n_0 times
  Let $\hat{f}_t^i$ be an estimator of f_i given t i.i.d. samples.
- 3-rd UCBs: Define, for any t such that $t > \tilde{d}(t)$, the following index function:

$$u_i(n,t) = \sup_{g \in \mathcal{F}} \left\{ s(g) : \mathbf{I}(\hat{f}_t^i, g) \leq \frac{\ln n}{t - \tilde{d}} \right\}$$

- For $n \geq n_0 \times N$, sample from bandit

$$\pi^*(n+1) = \arg\max_i \left\{ u_i^{CK}\left(n, T_{\pi^*}^i(n)\right) \right\}$$

breaking ties uniformly at random

- Then

$$\lim_n \frac{\mathbb{E}\left[T_{\pi^{CK}}^i(n)\right]}{\ln n} = \frac{1}{\mathbb{K}_i(s^*)}, \ \forall \ non-optimal \ i$$

π^* is a pure index policy

Asympt. Optimal UCB Policies for Normal Populations: X_k^i are iid $N(\mu_i, \sigma_i^2)$

Lai and Robbins (1985): Let $a_{nk} > 0$ ($n = 1, 2, \ldots,$ $k = 1, \ldots, n$) be sequences constants such that:

- for every fixed i such that a_{nk} is non-decreasing in $n \geq k$
- and there exist $\epsilon_n \to 0$ such that

$$|a_{nk} - \ln n/k| \leq \epsilon_n (\ln n/k)^{1/2} \;\forall\; k \leq n$$

Estimates $\hat{\mu}_i(k) = \hat{\theta}_i(k) = \sum_{m=1}^{k} X_m^i/k$ define

$$g_{nk}^i = g_{nk}^i(\hat{\mu}_i(k), a_{nk}) = \hat{\mu}_i(k) + \sigma(2a_{nk})^{1/2}$$

Final Comments: Past and Current Work - Regret

Asympt. Optimal UCB Policies for Normal Populations: X_k^i are iid $N(\mu_i, \sigma_i^2)$

Lai and Robbins (1985): Let $a_{nk} > 0$ $(n = 1, 2, \ldots,\ k = 1, \ldots, n)$ be sequences constants such that:

- for every fixed i such that a_{nk} is non-decreasing in $n \geq k$
- and there exist $\epsilon_n \to 0$ such that

$$|a_{nk} - \ln n/k| \leq \epsilon_n (\ln n/k)^{1/2}\ \forall\ k \leq n$$

Estimates $\hat{\mu}_i(k) = \hat{\theta}_i(k) = \sum_{m=1}^{k} X_m^i / k$ define

$$g_{nk}^i = g_{nk}^i(\hat{\mu}_i(k), a_{nk}) = \hat{\mu}_i(k) + \sigma(2a_{nk})^{1/2}$$

For $n+1 > N$ compute: j: $n+1 = mN + j$ and j_n^*:

$$\hat{\mu}_{j_n^*} = \max\{\hat{\mu}_i(n)\ :\ T^i_{\pi^{LR}}(n) > \delta n\}$$

$$\pi^{LR}(n+1) = \begin{cases} j & \text{if } \hat{\mu}_{j_n^*} < g_{nk}^i(\hat{\mu}_i(k), a_{nk}) \\ j_n^* & \text{otherwise} \end{cases}$$

Final Comments: Past and Current Work - Regret

Asympt. Optimal UCB Policies for Normal Populations: X_k^i are iid $N(\mu_i, \sigma_i^2)$

Lai and Robbins (1985): Let $a_{nk} > 0$ $(n = 1, 2, \ldots,\ k = 1, \ldots, n)$ be sequences constants such that:

- for every fixed i such that a_{nk} is non-decreasing in $n \geq k$
- and there exist $\epsilon_n \to 0$ such that

$$|a_{nk} - \ln n/k| \leq \epsilon_n (\ln n/k)^{1/2} \ \forall \ k \leq n$$

Estimates $\hat{\mu}_i(k) = \hat{\theta}_i(k) = \sum_{m=1}^{k} X_m^i / k$ define

$$g_{nk}^i = g_{nk}^i(\hat{\mu}_i(k), a_{nk}) = \hat{\mu}_i(k) + \sigma(2a_{nk})^{1/2}$$

For $n + 1 > N$ compute: j: $n + 1 = mN + j$ and j_n^*:

$$\hat{\mu}_{j_n^*} = \max\{\hat{\mu}_i(n) \ : \ T_{\pi^{LR}}^i(n) > \delta n\}$$

$$\pi^{LR}(n+1) = \begin{cases} j & \text{if } \hat{\mu}_{j_n^*} < g_{nk}^i(\hat{\mu}_i(k), a_{nk}) \\ j_n^* & \text{otherwise} \end{cases}$$

where $k = T_\pi^i(n)$ in the above.

Final Comments: Past and Current Work Continued

Asympt. Optimal UCB Policies for Normal Populations: X_k^i are iid $N(\mu_i, \sigma_i^2)$

Katehakis and Robbins (1995): μ_i unknown **and** σ_i^2 **known**
A policy π_g that first samples each bandit once, then for $n \geq N+1$,

$$u_i^{KR}(\hat{\mu}_i(n), n) = \overline{X}^i_{T^i_\pi(n)} + \sigma_i(\, 2\, log\, n / T^i(n))^{1/2}$$

Burnetas and Katehakis (1996): both μ_i **and** σ_i^2 unknown
A policy π_g that first samples each bandit once, then for $t \geq N+1$,

$$u_i^{BK}(\underline{\hat{\theta}}_i(n), n) = \overline{X}^i_{T^i_\pi(n)} + \hat{\sigma}_i(T^i(n)\,)(n^{2/(T^i(n))} - 1)^{1/2}$$

$O(\ln(n))$ regret open problem in 1996

Cowan, Honda and Katehakis (2015): both μ_i **and** σ_i^2 unknown
A policy π_g that first samples each bandit once, then for $t \geq N+1$,

$$u_i^{CHK}(\underline{\hat{\theta}}_i(n), n) = \overline{X}^i_{T^i_\pi(n)} + \hat{\sigma}_i(T^i(n)\,)(n^{2/(T^i(n)-2)} - 1)^{1/2}$$

Existence of UF policy π_{ACF}:

Policy π_{ACF} (UCB1-NORMAL). At each $n = 1, 2, \ldots$:

i) Sample from any bandit i for which $T^i_{\pi_{\text{ACF}}}(n) < \lceil 8 \ln n \rceil$.

ii) If $T^i_{\pi_{\text{ACF}}}(n) > \lceil 8 \ln n \rceil$, for all $i = 1, \ldots, N$, sample from bandit $\pi_{\text{ACF}}(n+1)$ with

$$\pi_{\text{ACF}}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + \hat{\sigma}_i(T^i(n))(16 \, log \, n / T^i(n))^{1/2} \right\}.$$

And the bound

$$R_{\pi_{\text{ACF}}}(n) \leq M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) \ln n + C_{\text{ACF}}(\underline{\mu}), \ \forall n \ and \ \forall (\underline{\mu}, \underline{\sigma}^2)$$

with

$$M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) = 256 \sum_{i:\mu_i \neq \mu^*} \frac{\sigma_i^2}{\Delta_i} + 8 \sum_{i=1}^N \Delta_i ,$$

$$C_{\text{ACF}}(\underline{\mu}) = (1 + \frac{\pi^2}{2}) \sum_{i=1}^N \Delta_i .$$

Existence of UF policy π_{ACF}:

Policy π_{ACF} (UCB1-NORMAL). At each $n = 1, 2, \ldots$:

i) Sample from any bandit i for which $T^i_{\pi_{\text{ACF}}}(n) < \lceil 8 \ln n \rceil$.

ii) If $T^i_{\pi_{\text{ACF}}}(n) > \lceil 8 \ln n \rceil$, for all $i = 1, \ldots, N$, sample from bandit $\pi_{\text{ACF}}(n+1)$ with

$$\pi_{\text{ACF}}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + \hat{\sigma}_i(T^i(n))(16 \, log \, n / T^i(n))^{1/2} \right\}.$$

And the bound

$$R_{\pi_{\text{ACF}}}(n) \leq M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) \ln n + C_{\text{ACF}}(\underline{\mu}), \ \forall n \ and \ \forall (\underline{\mu}, \underline{\sigma}^2)$$

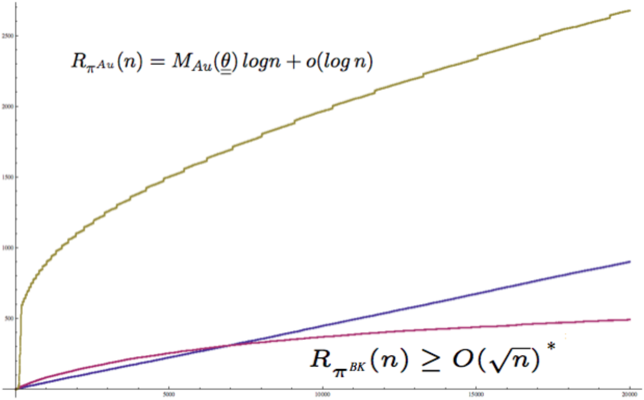
with

$$M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) = 256 \sum_{i: \mu_i \neq \mu^*} \frac{\sigma_i^2}{\Delta_i} + 8 \sum_{i=1}^{N} \Delta_i,$$

$$C_{\text{ACF}}(\underline{\mu}) = (1 + \frac{\pi^2}{2}) \sum_{i=1}^{N} \Delta_i.$$

$R_{\pi_{\text{ACF}}}(n) \leq M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) \ln n + o(\ln n)$. $\ln n = o(n^\alpha)$ for all $\alpha > 0$ and $R_{\pi_{\text{ACF}}}(n) \geq 0$,
i.e., π_{ACF} is uniformly fast convergent.

Regret Comparison



Π_1 : $X_{11}, X_{12}, \ldots$ **iid** $N(8.1, 1)$

Π_2 : $X_{21}, X_{22}, \ldots$ **iid** $N(8.1, 4)$

Π_3 : $X_{31}, X_{32}, \ldots$ **iid** $N(7.9, 0.5)$

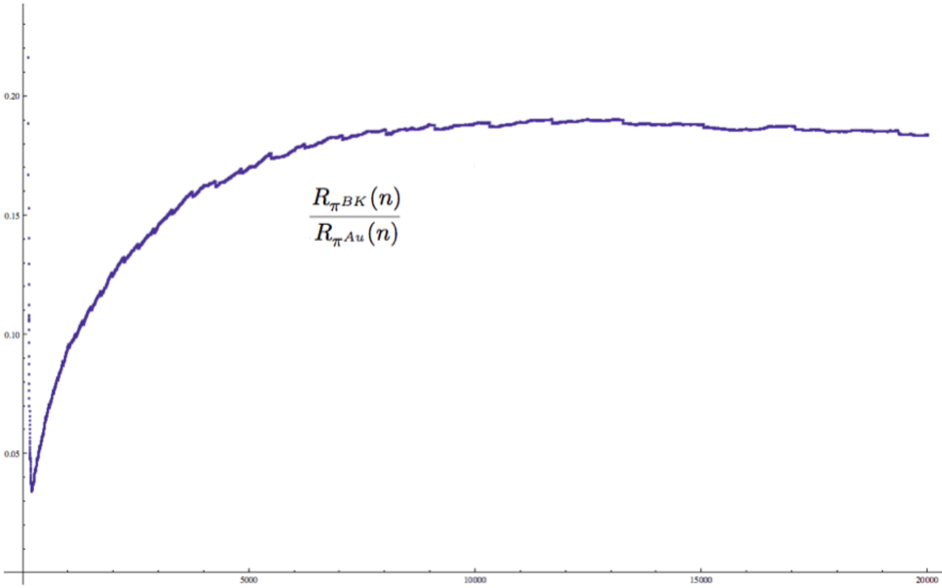
Π_4 : $X_{41}, X_{42}, \ldots$ **iid** $N(7, 3)$

Π_5 : $X_{51}, X_{52}, \ldots$ **iid** $N(-1, 1)$

Π_6 : $X_{61}, X_{62}, \ldots$ **iid** $N(0, 4)$

*: Cowan et al (2015)

Regret Comparison - Continued



- Unknown Variance: π^* an index policy based on $u_i^{CHK}(n)$

$$u_i^{CHK}(n) = \bar{X}^i_{T^i(n)} + \hat{\sigma}^i_{T^i(n)} \sqrt{n^{\frac{2}{T^i(n)-2}} - 1}$$

Cowan et al (2015)

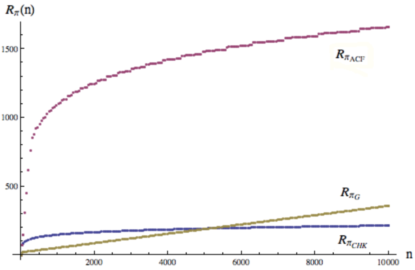
$$u_i^{BK}(n) = \bar{X}^i_{T^i(n)} + \hat{\sigma}^i_{T^i(n)} \sqrt{n^{\frac{2}{T^i(n)}} - 1}$$

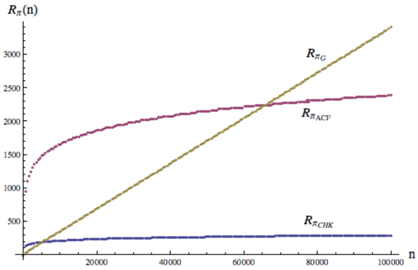
Burnetas and Katehakis (1996)

Results:

$$\lim_n \frac{R_{\pi_{\mathsf{CHK}}}(n)}{\ln n} = M^{\mathsf{BK}}(\underline{\mu}, \underline{\sigma}^2)$$

$$R_{\pi_{\mathsf{CHK}}}(n) \leq \sum_{i:\mu_i \neq \mu^*} \left(\frac{2 \ln n}{\ln\left(1 + \frac{\Delta_i^2}{\sigma_i^2} \frac{(1-\epsilon)^2}{(1+\epsilon)}\right)} + \sqrt{\frac{\pi}{2e}} \frac{8\sigma_i^3}{\Delta_i^3 \epsilon^3} \ln \ln n + \frac{8}{\epsilon^2} + \frac{8\sigma_i^2}{\Delta_i^2 \epsilon^2} + 4 \right) \Delta_i.$$





Figures 1 & 2 show the results of a small simulation study, implementing policies $\pi_{\mathrm{CHK}}, \pi_{\mathrm{ACF}}$, and π_G a 'greedy' policy that always activates the bandit with the current highest average. Simulation was done with six populations, with means and variances given in the table below.

μ_i	8	8	7.9	7	-1	0
σ_i^2	1	1.4	0.5	3	1	4

Each policy was implemented over a horizon of 10,000 and 100,000 activations, each replicated 10,000 times to produce a good estimate of the average regret $R_\pi(n)$ over the times indicated.

$$R_{\pi_{CHK}}(n) = M^{BK}(\underline{\theta}) \ln n + o(\ln n)$$

$M^{BK}(\underline{\theta}) = \sum_{i \in \mathcal{B}(\underline{\theta})} 1/ \inf_{\underline{\theta}'_i \in \underline{\Theta}_i} \{\mathbb{I}(\underline{\theta}_i, \underline{\theta}'_i) : \mu(\underline{\theta}'_i) > \mu(\underline{\theta}^*)\}$

Bounds and Limits:

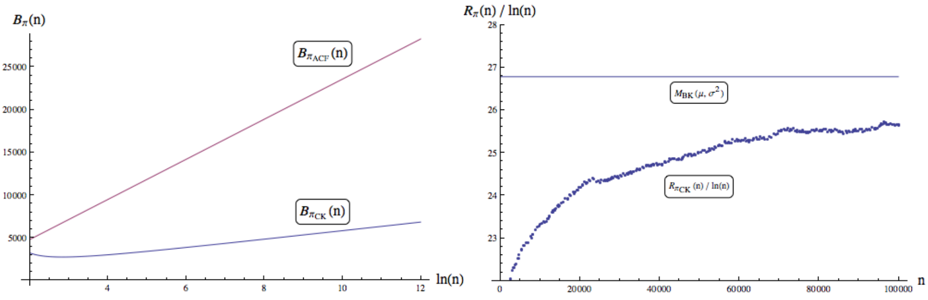


Figure: Left: Plots of $B_{\pi_{\mathsf{ACF}}}(n)$ and $B_{\pi_{\mathsf{CHK}}}(n)$. Right: Convergence of $R_{\pi_{\mathsf{CHK}}}(n)/\ln(n)$ to $M^{\mathsf{BK}}(\underline{\mu}, \underline{\sigma}^2)$

Figure 2 shows first (left) a comparison of the theoretical bounds on the regret, $B_{\pi_{\mathsf{ACF}}}(n)$ and $B_{\pi_{\mathsf{CHK}}}(n)$ representing their theoretical regret bounds respectively, for the means and variances indicated in the table below. Additionally, Figure 2 (right) shows the convergence of $R_{\pi_{\mathsf{CHK}}}(n)/\ln n$ to the theoretical lower bound $M^{\mathsf{BK}}(\underline{\mu}, \underline{\sigma}^2)$.
To produce a good estimate of the average regret $R_\pi(n)$ over the times indicated, each policy was implemented over a horizon of 100,000 activations, each replicated 10,000 times.

Policy π_{TS} (TS-NORMAL$^{\alpha}$)

i) Initially, sample each bandit $\tilde{n} \geq \max(2, 3 - \lfloor 2\alpha \rfloor)$ times.

ii) For $n \geq \tilde{n}:$ For each i generate a random sample U_n^i from a posterior distribution for μ_i, given $\left(\bar{X}^i_{T^i_\pi(n)}, \hat{\sigma}_i^2(T^i_\pi(n))\right)$, and a prior for $\left(\mu_i, \sigma_i^2\right) \propto \left(\sigma_i^2\right)^{-1-\alpha}$.

iii) Then, take

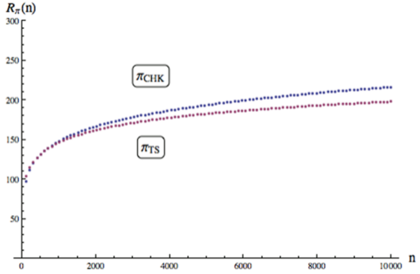
$$\pi_{\mathsf{TS}}(n+1) = \arg\max_i U_n^i.$$

$$\lim_n \frac{R_{\pi_{TS}}(n)}{\ln n} = M^{\mathsf{BK}}(\underline{\mu}, \underline{\sigma}^2), \quad \forall (\underline{\mu}, \underline{\sigma}^2) \qquad !$$

Thompson Sampling - Normal with unknown μ_i & σ_i^2

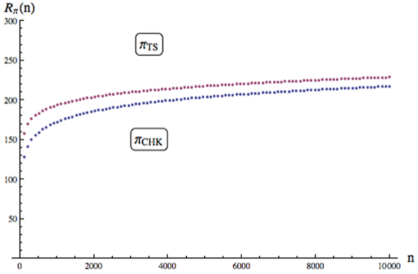
Cowan, Honda and Katehakis 2015

Numerical Regret Comparison of π_{CHK} and π_{TS}



μ_i	8	8	7.9	7	-1	0
σ_i^2	1	1.4	0.5	3	1	4

Table 1



μ_i	10	9	8	7	-1	0
σ_i^2	8	1	1	0.5	1	4

Table 2

$R_{\pi_{CHK}}(n)$ and $R_{\pi_{TS}}(n)$ for the parameters, of Table 1, left and Table 2, right.

For each i, $f_i \in \mathcal{F}$ Uniform on $[a_i, b_i]$

$$\mu(f_i) = (a_i + b_i)/2$$

$$\hat{a}^i_t = \min_{t' \leq t} X^i_{t'} \quad \& \quad \hat{b}^i_t = \max_{t' \leq t} X^i_{t'}$$

$$\boxed{u^i_{CK}(n, t, \hat{f}^i_t) = \hat{a}^i_t + \frac{1}{2}\left(\hat{b}^i_t - \hat{a}^i_t\right) n^{\frac{1}{t-2}} \quad \textit{asymptotically optimal!}}$$

$$\pi_{\mathsf{CK}}(n+1) = \arg\max_i \; u^i_{CK}(n, t, \hat{f}^i_t)$$

$$M^{\mathrm{BK}}(\{(a_i, b_i)\}) = \sum_{i:\mu_i \neq \mu^*} \frac{\Delta_i}{\ln\left(1 + \frac{2\Delta_i}{b_i - a_i}\right)}$$

For each i, $f_i \in \mathcal{F}$ Uniform on $[a_i, b_i]$

$$\mu(f_i) = (a_i + b_i)/2$$

$$\hat{a}^i_t = \min_{t' \leq t} X^i_{t'} \quad \& \quad \hat{b}^i_t = \max_{t' \leq t} X^i_{t'}$$

$$u^i_{CK}(n, t, \hat{f}^i_t) = \hat{a}^i_t + \frac{1}{2}\left(\hat{b}^i_t - \hat{a}^i_t\right) n^{\frac{1}{t-2}} \quad \textit{asymptotically optimal!}$$

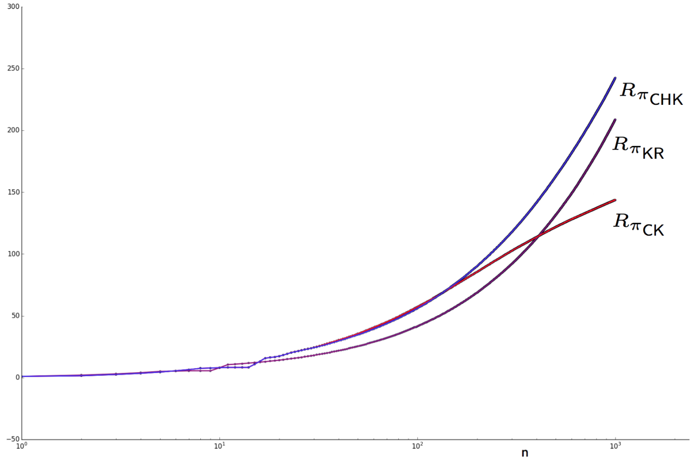
$$\pi_{\mathsf{CK}}(n+1) = \arg\max_i \; u^i_{CK}(n, t, \hat{f}^i_t)$$

$$M^{\mathrm{BK}}(\{(a_i, b_i)\}) = \sum_{i:\mu_i \neq \mu^*} \frac{\Delta_i}{\ln\left(1 + \frac{2\Delta_i}{b_i - a_i}\right)}$$

$$u^i_{BK}(n, t, \hat{f}^i_t) = \hat{a}^i_t + \frac{1}{2}\left(\hat{b}^i_t - \hat{a}^i_t\right) n^{1/t} \quad \textit{asymptotically optimal?}$$

Short Time Horizon: Numerical regret comparison of π_{CK}, π_{KR}, and π_{CHK}, for the 6 bandits with parameters given in Table 1.

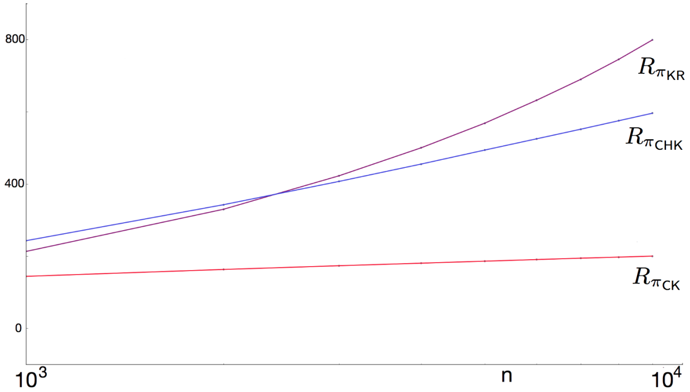
Average values over $20,000$ repetitions.



i	1	2	3	4	5	6
a_i	0	0	0	1	1	1
b_i	10	9	8	9.5	10	5

Longer Time Horizon: Numerical regret comparison of π_{CK}, π_{KR}, and π_{CHK}, for the 6 bandits with parameters given in Table 1.

Average values over $10,000$ repetitions.



π_{KR}	$u^i_{\mathsf{KR}}(n,t) = \bar{X}^i_t + \hat{\sigma}^i(t)\sqrt{\frac{2\ln n}{t}}$
π_{CHK}	$u^i_{\mathsf{CHK}}(n,t) = \bar{X}^i_t + \hat{\sigma}^i(t)\sqrt{n^{\frac{2}{t-2}} - 1}$
π_{CK}	$u^i_{CK}(n,t,\hat{f}^i_t) = \hat{a}^i_t + \frac{1}{2}\left(\hat{b}^i_t - \hat{a}^i_t\right) n^{\frac{1}{t-2}}$

References

- Auer, P.; Cesa-Bianchi, N. and P. Fischer (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235 ? 256, .
- Burnetas, A. N., and Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics* 17(2):122-142.
- Cappé, O.; Garivier, A.; Maillard, O.-A.; Munos, R.; and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics* 41(3):1516-1541.
- Cowan, W. and Katehakis, M. N. (2015a). Asymptotically Optimal Sequential Experimentation Under Generalized Ranking, arXiv:1510.02041
- Cowan, W.; Honda, J.; and Katehakis, M. N. (2015). Asymptotic optimality, finite horizon regret bounds, and a solution to an open problem. *Journal of Machine Learning Research, to appear; preprint arXiv:1504.05823.*
- Cowan, W. and Katehakis, M. N. (2015b). An Asymptotically Optimal Policy for Uniform Bandits of Unknown Support, arXiv:1505.01918
- Cowan, W. and Katehakis, M. N. (2015c). Asymptotic Behavior of Minimal-Exploration Allocation Policies: Almost Sure, Arbitrarily Slow Growing Regret, arXiv:1505.02865
- Honda J., and Takemura, A. (2010). An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, 67-69. Cities.
- Honda J., and Takemura, A. (2011). An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning* 85(3)361-391.
- Honda J., and Takemura, A. (2013). Optimality of Thompson sampling for Gaussian bandits depends on priors. *arXiv preprint arXiv:1311.1894.*
- Katehakis M.N. and H.E. Robbins (1995). Sequential choice from several populations, *Proceedings of the National Academy of Sciences USA*, 92, 8584 -8565.
- Kaufmann Emilie (2015). Analyse de strategies Bayesiennes et frequentistes pour l'allocation sequentielle de ressources. Doctorat, ParisTech., Jul. 31 2015.
- Lai, T. L., and Robbins, H.E. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics* 6(1):4-22.
- May, B. C.; Korda, N.; Lee, A. and Leslie, D. S. (2012). Optimistic Bayesian sampling in contextual-bandit problems, *The Journal of Machine Learning Research*, 13, 2069–2106
- Robbins, H.E. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Monthly* 58:527-536.