Multi-armed bandit (MAB) problems, models and algorithms

Michael N. Katehakis

Rutgers University

June 27, 2015

Talk based on joint work with C. Wesley Cowan, Rutgers University
Outline

- Exploration vs Exploitation: The MAB problem
- Motivation with Applications
- Thompson 1933 and Robbins 1952
- Gittins Formulation
- Lai-Robbins Formulation
- Good Fast and not so good UCB policies
- Current work
- Questions
Beginnings: Thomson 1933 - Robbins 1952

Example

\[ \Pi_1 \]

\[ \Pi_2 \]

\[ X_{11}, X_{12}, X_{13}, \ldots \]

\[ X_{21}, X_{22}, X_{23}, \ldots \]

\[ X_{it} = \begin{cases} 
1 & \text{patient recovers with probability } p_i \\
-1 & \text{patient dies with probability } 1 - p_i 
\end{cases} \]

Traditional Allocation

Assign 100 to \( \Pi_1 \) observe \( 1, -1, -1, -1, 1, \ldots, -1, 1 \) with \( \sum_{1}^{100} = 10 - 90 \Rightarrow \hat{p}_1 = \frac{10}{100} \)

Assign 100 to \( \Pi_2 \) observe \( 1, 1, -1, 1, 1, \ldots, 1 \) with \( \sum_{1}^{100} = 99 - 1 \Rightarrow \hat{p}_2 = \frac{99}{100} \)

``Killed'' 90 on \( \Pi_1 \) to learn

CAN WE DO BETTER?
Example

\[ \Pi_1, \Pi_2 \]

\[ X_{11}, X_{12}, X_{13}, \ldots \quad X_{21}, X_{22}, X_{23}, \ldots \]

\[ \begin{align*}
X_{it} &= \begin{cases} 
1 & \text{patient recovers with probability } p_i \\
-1 & \text{patient dies with probability } 1 - p_i
\end{cases}
\end{align*} \]

Traditional Allocation

Assign 100 to \( \Pi_1 \) observe 1,-1,-1,-1,-1, , , , , -1,1 with \( \sum_{1}^{100} = 10 - 90 \rightarrow \hat{p}_1 = \frac{10}{100} \)

Assign 100 to \( \Pi_2 \) observe 1,1,-1,1,1, , , , , 1 with \( \sum_{1}^{100} = 99 - 1 \rightarrow \hat{p}_2 = \frac{99}{100} \)

``Killed” 90 on \( \Pi_1 \) to learn

CAN WE DO BETTER?

Given 2 populations (treatments) that generate 0, 1 outcomes:

\[ \Pi_1 : \quad X_1^1, X_2^1, \ldots \quad p_1 = Pr\{X_k^1 = 1\} \quad p_1 \sim f_0^1(p) \]

\[ \Pi_2 : \quad X_1^2, X_2^2, \ldots \quad p_2 = Pr\{X_k^2 = 1\} \quad p_2 \sim f_0^2(p) \]
Given 2 populations (treatments) that generate 0, 1 outcomes:

\[ \Pi_1 : \quad X_1^1, X_2^1, \ldots \quad p_1 = Pr\{X_k^1 = 1\} \quad p_1 \sim f_0^1(p) \]

\[ \Pi_2 : \quad X_1^2, X_2^2, \ldots \quad p_2 = Pr\{X_k^2 = 1\} \quad p_2 \sim f_0^2(p) \]

For \( t = 0, \ldots \) and an adaptive policy \( \pi \) let:
\( \pi(t) = i \) indicate the event that population \( i = 1, 2 \) is sampled at time \( t \), and let

\[ T^i_\pi(n) = \sum_{t=1}^{n} \mathbf{1}_{\pi(t) = i} \]
Given 2 populations (treatments) that generate 0, 1 outcomes:

\[ \Pi_1 : \quad X^1_1, X^1_2, \ldots \quad p_1 = Pr\{X^1_k = 1\} \quad p_1 \sim f^1_0(p) \]
\[ \Pi_2 : \quad X^2_1, X^2_2, \ldots \quad p_2 = Pr\{X^2_k = 1\} \quad p_2 \sim f^2_0(p) \]

For \( t = 0, \ldots \) and an adaptive policy \( \pi \) let:
\( \pi(t) = i \) indicate the event that population \( i = 1, 2 \) is sampled at time \( t \), and let

\[ T^i_\pi(n) = \sum_{t=1}^{n} 1_{\pi(t)=i} \]

It is convenient to define \( T^i_\pi(0) = 0 \) for all \( i, \pi \).

\[ Data(n) = [(X_{11}, X_{12}, \ldots, X_{1,T^1_\pi(n)}), (X_{21}, X_{22}, \ldots, X_{2,T^2_\pi(n)})] \]

\[ \pi^*(t) = \begin{cases} 1 & \text{with } Pr\{p_1 > p_2|Data(n)\} \\ 2 & \text{with } Pr\{p_1 < p_2|Data(n)\} \end{cases} \]
Given 2 populations (treatments) that generate outcomes:

\[ \Pi_i : \quad X^i_1, X^i_2, \ldots \quad \text{with} \quad \mu_i = \mathbb{E}[X^i_k] = \int_{-\infty}^{+\infty} x \, dF^i(x) < \infty \text{ unknown} \]

\[ S^\pi(n) = \sum_{i=1}^{N} \sum_{k=1}^{T^i_\pi(n)} X^i_k \text{ (max)} \]

Defined measures of regret:

\[ R'_\pi(n) = n\mu^* - S^\pi(n) = n\mu^* - \sum_{i=1}^{N} \sum_{k=1}^{T^i_\pi(n)} X^i_k, \]

\[ R_\pi(n) = n\mu^* - \mathbb{E}S^\pi(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[T^i_\pi(n)\right]. \]

\[ \mu^* = \max\{\mu_1, \mu_2\} \quad \text{and} \quad \Delta_i = \mu^* - \mu_i \]
Given 2 populations (treatments) that generate outcomes:

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots \text{ with } \mu_i = \mathbb{E}[X_k^i] = \int_{-\infty}^{+\infty} x dF_i(x) < \infty \text{ unknown} \]

\[ S_{\pi}(n) = \sum_{i=1}^{N} \sum_{k=1}^{T^i_\pi(n)} X_k^i \text{ (max)} \]

Defined measures of regret:

\[ R'_{\pi}(n) = n \mu^* - S_{\pi}(n) = n \mu^* - \sum_{i=1}^{N} \sum_{k=1}^{T^i_\pi(n)} X_k^i, \]

\[ R_{\pi}(n) = n \mu^* - \mathbb{E}S_{\pi}(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[T^i_\pi(n)\right]. \]

\[ \mu^* = \max\{\mu_1, \mu_2\} \text{ and } \Delta_i = \mu^* - \mu_i \]

Constructed a modified (outside two sparse sequences of forced choices) ‘play the winner’ (greedy) policy, \( \pi_R \), such that with probability one, as \( n \to \infty \),

\[ S_{\pi_R}(n)/n \to \mu^*, \text{ i.e.,} \]

\[ R'_{\pi_R}(n) = o(n) \text{ (a.s.), as } n \to \infty. \]
Given $N \geq 2$ populations (treatments) that generate outcomes:

$\Pi_i : X_{i1}, X_{i2}, \ldots$  

_Stationary Process with known Probability Law_
The MAB Problem

Given $N \geq 2$ populations (treatments) that generate outcomes:

$\Pi_i : X_{i1}, X_{i2}, \ldots$  St. Proc. with known Probability Law

$S^i_\pi(t)$ to be the round at which process $i$ is activated for the $t^{th}$ time (and $X^i(t)$ is collected), when the controller operates according to policy $\pi$:

$$S^i_\pi(0) = \inf\{s \geq 0 : \pi(s) = i\},$$

$$S^i_\pi(t + 1) = \inf\{s > S^i_\pi(t) : \pi(s) = i\}.$$

$$v_\pi = \sum_{i=1}^{N} \mathbb{E} \sum_{k=0}^{\infty} \beta^{S^i_\pi(k)} X^i_k \quad \text{(max)} \quad \text{some } \beta \in (0, 1)$$
The MAB Problem

Given $N \geq 2$ populations (treatments) that generate outcomes: $\Pi_i: X_{i1}, X_{i2}, \ldots$  

St. Proc. with known Probability Law $S^i_\pi(t)$ to be the round at which process $i$ is activated for the $t^{th}$ time (and $X^i(t)$ is collected), when the controller operates according to policy $\pi$:

$$S^i_\pi(0) = \inf \{ s \geq 0 : \pi(s) = i \},$$
$$S^i_\pi(t + 1) = \inf \{ s > S^i_\pi(t) : \pi(s) = i \}.$$  

$$v_\pi = \sum_{i=1}^{N} \sum_{k=0}^{\infty} \beta^{S^i_\pi(k)} X^i_k \quad \text{(max)} \quad \text{some } \beta \in (0, 1)$$

$$\pi^*(n|\text{Data}(n)) = \arg \max \{ \gamma_i(X^i_{T^i_{\pi^*}(n)}) \}$$

Gittins DAI index:

$$\gamma_i(X^i_{T^i_{\pi^*}(n)}) = \sup_{\tau > T^i_{\pi^*}(n)} \frac{\mathbb{E} \sum_{k=0}^{\tau-1} \beta^k X^i_k | X^i_{T^i_{\pi^*}(n)}}{\mathbb{E} \sum_{k=0}^{\tau-1} \beta^k X^i_k | X^i_{T^i_{\pi^*}(n)}}

= (1 - \beta) m_i(X^i_{T^i_{\pi^*}(n)})$$
Given \( N \geq 2 \) populations (treatments) that generate outcomes:
\[ \Pi_i : \quad X_{i1}, X_{i2}, \ldots \quad \text{St. Proc. with known Probability Law} \]
\( S^i_\pi(t) \) to be the round at which process \( i \) is activated for the \( t^{th} \) time (and \( X^i(t) \) is collected), when the controller operates according to policy \( \pi \):
\[
S^i_\pi(0) = \inf\{s \geq 0 : \pi(s) = i\},
\]
\[
S^i_\pi(t + 1) = \inf\{s > S^i_\pi(t) : \pi(s) = i\}.
\]

\[
u_\pi = \sum_{i=1}^{N} \mathbb{E} \sum_{k=0}^{\infty} \beta^{S^i_\pi(k)} X^i_k \quad \text{(max)} \quad \text{some } \beta \in (0, 1)
\]

\[
\pi^*(n|\text{Data}(n)) = \arg\max\{\gamma_i(X^i_{T^i_{\pi^*}(n)})\}
\]

**Gittins DAI index:**

\[
\gamma_i(X^i_{T^i_{\pi^*}(n)}) = \sup_{\tau > T^i_{\pi^*}(n)} \frac{\mathbb{E} \sum_{k=0}^{\tau-1} \beta^k X^i_k | X^i_{T^i_{\pi^*}(n)}}{\mathbb{E} \sum_{k=0}^{\tau-1} \beta^k | X^i_{T^i_{\pi^*}(n)}}
\]
\[
= (1 - \beta) m_i(X^i_{T^i_{\pi^*}(n)})
\]

\( m_i(\cdot) \) restart in state index - Katehakis and Veinott (1987)
The MAB Problem

Gittins 1979

Extentions

- El Karoui & Karatzas (1994): Continuous Time
- El Karoui & Karatzas (1993): Discrete Time
- Kaspi & Mandelbaum (1996): General treatment
- Cowan & Katehakis (2015): General Depreciation - Commitment Times
The MAB Problem  Lai - Robbins 1985

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots \quad f(x; \theta_i) \text{ unknown } \theta_i \in \Theta. \]

where \( f(\cdot; \cdot) \) is known univariate density w.r.t. some measure \( \nu \)

Let \( \theta = (\theta_1, \ldots, \theta_N) \)

\[ \mu_i = \mu(\theta_i) = \mathbb{E}X_1^i, \quad \mu^* = \mu(\theta^*_i), \quad \Delta_i(\theta_i) = \mu^* - \mu(\theta_i), \]

\[ \mathbb{I}(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, d\nu(x) \]

be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).
The MAB Problem  Lai - Robbins 1985

\( \Pi_i : \quad X^i_1, X^i_2, \ldots \ f(x; \theta_i) \) unknown \( \theta_i \in \Theta \).

where \( f(.;.) \) is known univariate density w.r.t. some measure \( \nu \)

Let \( \theta = (\theta_1, \ldots, \theta_N) \)

\[ \mu_i = \mu(\theta_i) = \mathbb{E}X^i_1, \mu^* = \mu(\theta^*_i), \Delta_i(\theta_i) = \mu^* - \mu(\theta_i), \]

\[ \mathbb{I}(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, d\nu(x) \] be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).

Recall:

\[ R_\pi(n) = R_\pi(n|\theta) = n\mu^* - \mathbb{E}S_\pi(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E}[T^i_\pi(n)] \]

Types of Policies

- **UC u-consistent** \( \pi : \) \( R_\pi(n|\theta) = o(n) \) (as \( n \to \infty \)) \( \forall \theta \)

- **UF u-fast** \( \pi : \) \( R_\pi(n|\theta) = O(\log n) = M(\theta) \log n + o(\log n) = o(n^\alpha), \forall \alpha > 0 \)
\[\Pi_i : \quad X^i_1, X^i_2, \ldots \quad f(x; \theta_i) \text{ unknown } \theta_i \in \Theta.\]

where \(f(.; .)\) is known univariate density w.r.t. some measure \(\nu\)
Let \(\theta = (\theta_1, \ldots, \theta_N)\)
\(\mu_i = \mu(\theta_i) = \mathbb{E}X^i_1, \mu^* = \mu(\theta_{i^*}), \Delta_i(\theta_i) = \mu^* - \mu(\theta_i),\)
\(\mathbb{I}(\theta, \theta') = \int^{-\infty}_\infty \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, dv(x)\) be K-L divergence between \(f(x; \theta)\) and \(f(x; \theta')\).
Recall:
\[R_\pi(n) = R_\pi(n|\theta) = n \mu^* - \mathbb{E}S_\pi(n) = \sum_{i=1}^N \Delta_i \mathbb{E} [T^i_\pi(n)]\]

**Types of Policies**

- **UC** u-consistent \(\pi\): \(R_\pi(n|\theta) = o(n) \text{ (as } n \to \infty) \quad \forall \theta\)
- **UF** u-fast \(\pi\): \(R_\pi(n|\theta) = O(\log n) = M(\theta) \log n + o(\log n) = o(n^\alpha), \quad \forall \alpha > 0\)

How good can ‘fast’ policies be?
\[ \Pi_i : \quad X_1^i, X_2^i, \ldots \quad f(x; \theta_i) \text{ unknown } \theta_i \in \Theta. \]

where \( f(\cdot, \cdot) \) is known univariate density w.r.t. some measure \( \nu \)
Let \( \underline{\theta} = (\theta_1, \ldots, \theta_N) \)
\( \mu_i = \mu(\theta_i) = \mathbb{E} X_1^i, \mu^* = \mu(\theta_{i^*}), \Delta_i(\theta_i) = \mu^* - \mu(\theta_i), \)
\( \Pi(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, dv(x) \) be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).
Recall:
\[ R_\pi(n) = R_\pi(n|\theta) = n\mu^* - \mathbb{E} S_\pi(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E} \left[ T_i^\pi(n) \right] \]

Types of Policies

- **UC** \( u \)-consistent \( \pi \): \( R_\pi(n|\theta) = o(n) \) (as \( n \to \infty \)) \( \forall \theta \)
- **UF** \( u \)-fast \( \pi \): \( R_\pi(n|\theta) = \mathcal{O}(\log n) = M(\underline{\theta}) \log n + o(\log n) = o(n^\alpha), \forall \alpha > 0 \)

How good can ‘fast’ policies be?

They showed, under mild regularity conditions its regret satisfies
\[
\lim_{n} \frac{R_\pi(n)}{\ln n} \geq M^{LR}(\theta), \quad \text{for all } \theta_i \in \Theta, \text{ and } i = 1, \ldots, N,
\]
where
\[
M^{LR}(\theta) = \sum_{i: \mu(\theta_i) \neq \mu^*} \Delta_i(\theta_i) / \Pi(\theta_i, \theta^*)
\]
\[ \Pi_i : \quad X^i_1, X^i_2, \ldots \quad f(x; \theta_i) \text{ unknown } \theta_i \in \Theta. \]

where \( f(.; .) \) is known univariate density w.r.t. some measure \( \nu \)

Let \( \underline{\theta} = (\theta_1, \ldots, \theta_N) \)

\[
\mu_i = \mu(\theta_i) = \mathbb{E}X^i_1, \quad \mu^* = \mu(\theta^*_i), \quad \Delta_i(\theta_i) = \mu^* - \mu(\theta_i),
\]

\[
\mathbb{I}(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, d\nu(x)
\]

be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).

Recall:

\[
R_\pi(n) = R_\pi(n|\theta) = n\mu^* - \mathbb{E}S_\pi(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E}[T^i_\pi(n)]
\]

Types of Policies

- **UC** \( u \)-consistent \( \pi \): \( R_\pi(n|\theta) = o(n) \) (as \( n \to \infty \)) \( \forall \theta \)

- **UF** \( u \)-fast \( \pi \): \( R_\pi(n|\theta) = O(\log n) = M(\theta) \log n + o(\log n) = o(n^\alpha), \ \forall \alpha > 0 \)

How good can ‘fast’ policies be?

They showed, under mild regularity conditions its regret satisfies

\[
\lim_{n} \frac{R_\pi(n)}{\ln n} \geq M^{LR}(\theta), \quad \text{for all } \theta_i \in \Theta, \text{ and } i = 1, \ldots, N,
\]

where

\[
M^{LR}(\theta) = \sum_{i: \mu(\theta_i) \neq \mu^*} \Delta_i(\theta_i)/\mathbb{I}(\theta_i, \theta^*)
\]

- **UM** optimal \( \pi \):

\[
\lim_{n} \frac{R_\pi(n)}{\ln n} = M^{LR}(\theta), \quad \text{for all } \theta_i \in \Theta, \text{ and } i = 1, \ldots, N
\]
The MAB Problem

Burnetas - Katehakis 1996

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots f_i(x; \theta_i) \quad \text{unknown} \quad \theta_i \in \Theta_i \]

where \( f_i(.;.) \) is known univariate density w.r.t. some measure \( \nu \)

- Let \( \Theta = (\theta_1, \ldots, \theta_N) \)
- \( \mu_i = \mu(\theta_i) = \mathbb{E}X_1^i, \mu^* = \mu(\theta_{i*}), \Delta_i(\theta_i) = \mu^* - \mu(\theta_i) \)
- \( I(\theta, \theta') = \int_{-\infty}^{\infty} \ln \left( \frac{f(x; \theta)}{f(x; \theta')} \right) f(x; \theta) \, dv(x) \) be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).
- \( O(\Theta) := \{ i : \mu(\theta_i) = \mu(\theta_{i*}) \} \)
- \( B(\Theta) := \{ i : \mu(\theta_i) < \mu(\theta_{i*}) \land \exists \theta'_i \in \Theta_i : \mu(\theta'_i) > \mu(\theta_{i*}) \} \)
The MAB Problem

Burnetas - Katehakis 1996

$\Pi_i : X_1^i, X_2^i, \ldots f_i(x; \theta_i)$ unknown $\theta_i \in \Theta_i$

where $f_i(\cdot; \cdot)$ is known univariate density w.r.t. some measure $\nu$

- Let $\underline{\theta} = (\theta_1, \ldots, \theta_N)$
- $\mu_i = \mu(\underline{\theta}_i) = E X_1^i, \mu^* = \mu(\underline{\theta}_*)$, $\Delta_i(\theta_i) = \mu^* - \mu(\theta_i)$
- $\Pi(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, dv(x)$ be K-L divergence between $f(x; \theta)$ and $f(x; \theta')$.
- $O(\theta) := \{ i : \mu(\theta_i) = \mu(\theta_*) \}$
- $B(\theta) := \{ i : \mu(\theta_i) < \mu(\theta_*) \text{ and } \exists \theta'_i \in \Theta_i : \mu(\theta'_i) > \mu(\theta_*) \}$
The MAB Problem

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \]

where \( f_i(.; .) \) is known univariate density w.r.t. some measure \( \nu \)

- Let \( \underline{\theta} = (\theta_1, \ldots, \theta_N) \)
- \( \mu_i = \mu(\underline{\theta_i}) = \mathbb{E}X_1^i \) \( \mu^* = \mu(\underline{\theta^*}) \), \( \Delta_i(\underline{\theta_i}) = \mu^* - \mu(\underline{\theta_i}) \),
- \( \Pi(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, dv(x) \) be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).
- \( O(\underline{\theta}) := \{ i : \mu(\underline{\theta_i}) = \mu(\underline{\theta^*}) \} \)
- \( B(\underline{\theta}) := \{ i : \mu(\underline{\theta_i}) < \mu(\underline{\theta^*}) \text{ & } \exists \underline{\theta'_i} \in \Theta_i : \mu(\underline{\theta'_i}) > \mu(\underline{\theta^*}) \} \)

- **UF** implies:

\[
\lim_{n} \frac{R_\pi(n)}{\ln n} \geq M^{BK}(\underline{\theta}), \quad \text{for all } \theta_i \in \Theta_i, \text{ and } i = 1, \ldots, N,
\]

where

\[
M^{BK}(\underline{\theta}) = \sum_{i \in B(\underline{\theta})} \Delta_i(\underline{\theta_i}) / \inf_{\underline{\theta'_i}} \{ \Pi(\theta_i, \underline{\theta'_i}) : \mu(\underline{\theta'_i}) > \mu(\underline{\theta^*}) \}
\]
\( \Pi_i : X^i_1, X^i_2, \ldots f_i(x; \theta_i) \) unknown \( \theta_i \in \Theta_i \)

where \( f_i(\cdot; \cdot) \) is known univariate density w.r.t. some measure \( \nu \)
- Let \( \underline{\theta} = (\theta_1, \ldots, \theta_N) \)
- \( \mu_i = \mu(\underline{\theta_i}) = \mathbb{E}X^i_1, \mu^* = \mu(\theta_i^*), \Delta_i(\theta_i) = \mu^* - \mu(\theta_i) \)
- \( \Pi(\theta, \theta') = \int_\mathbb{R} \ln \frac{f(x; \theta)}{f(x; \theta')} f(x; \theta) \, dv(x) \) be K-L divergence between \( f(x; \theta) \) and \( f(x; \theta') \).
- \( O(\underline{\theta}) := \{i : \mu(\theta_i) = \mu(\theta^*)\} \)
- \( B(\underline{\theta}) := \{i : \mu(\theta_i) < \mu(\theta^*) \& \exists \theta'_i \in \Theta_i : \mu(\theta'_i) > \mu(\theta^*)\} \)

- **UF** implies:
  \[
  \lim_{n} \frac{R_\pi(n)}{\ln n} \geq M^{BK}(\underline{\theta}), \quad \text{for all } \theta_i \in \Theta_i, \text{ and } i = 1, \ldots, N,
  \]

  where
  \[
  M^{BK}(\underline{\theta}) = \sum_{i \in B(\underline{\theta})} \Delta_i(\theta_i) / \inf_{\theta'_i} \{\Pi(\theta_i, \theta'_i) : \mu(\theta'_i) > \mu(\theta^*)\}
  \]

- **UM** \( \pi^o \) index based on upper confidence bounds (UCBs):
  \[
  u^{BK}(\hat{\theta}_i) = \sup_{\theta'_i \in \Theta_i} \left\{ \mu(\theta'_i) : \Pi(\hat{\theta}_i, \theta'_i) < \frac{\ln n}{T^o_\pi(n)} \right\}
  \]

  \[
  \lim_{n} \frac{R_{\pi^o}(n)}{\ln n} = M^{BK}(\underline{\theta}), \quad \text{for all } \theta_i \in \Theta_i, \text{ and } i = 1, \ldots, N
  \]
The MAB Problem

Lai - Robbins 1985 & Burnetas - Katehakis 1996

\( \Pi_i : \quad X_1^i, X_2^i, \ldots \ f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \)

\[
M_{LR}(\theta) = \sum_{i : \mu(\theta_i) \neq \mu^*} \Delta_i(\theta_i) \frac{\Delta_i(\theta_i)}{\Pi(\theta_i, \theta^*)}
\]

\( \Pi_i : \quad X_1^i, X_2^i, \ldots \ f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \)
The MAB Problem  

Lai - Robbins 1985 & Burnetas - Katehakis 1996

\[ \Pi_i : X^i_1, X^i_2, \ldots \text{ } f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \]

\[ M^{LR}(\theta) = \sum_{i : \mu(\theta_i) \neq \mu^*} \frac{\Delta_i(\theta_i)}{\Pi(\theta_i, \theta^*)} \]

\[ \Pi_i : X^i_1, X^i_2, \ldots \text{ } f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \]

\[ M^{BK}(\theta) = \sum_{i \in B(\theta)} \inf_{\theta_i'} \left\{ \Pi(\theta_i, \theta_i') : \mu(\theta_i') > \mu(\theta^*) \right\} \]

\[ B(\theta) := \left\{ i : \mu(\theta_i) < \mu(\theta^*) \text{ and } \exists \theta_i' \in \Theta_i : \mu(\theta_i') > \mu(\theta^*) \right\} \]
The MAB Problem  

Lai - Robbins 1985 & Burnetas - Katehakis 1996

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots, f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \]

\[ M^{LR}(\theta) = \sum_{i: \mu(\theta_i) \neq \mu^*} \frac{\Delta_i(\theta_i)}{\mathbb{I}(\theta_i, \theta^*)} \]

\[ \Pi_i : \quad X_1^i, X_2^i, \ldots, f_i(x; \theta_i) \text{ unknown } \theta_i \in \Theta_i \]

\[ M^{BK}(\theta) = \sum_{i \in B(\theta)} \frac{\Delta_i(\theta_i)}{\inf_{\theta_i'} \{ \mathbb{I}(\theta_i, \theta_i') : \mu(\theta_i') > \mu(\theta^*) \}} \]

\[ B(\theta) := \{ i : \mu(\theta_i) < \mu(\theta^*) \text{ & } \exists \theta_i' \in \Theta_i : \mu(\theta_i') > \mu(\theta^*) \} \]

- **UM** \( \pi^o \) index based on UCBs:

\[ u_{i}^{BK}(\hat{\theta}_i) = \sup_{\theta_i' \in \Theta_i} \left\{ \mu(\theta_i') : \mathbb{I}(\hat{\theta}_i, \theta_i') < \frac{\ln n}{T_{\pi^o}^i(n)} \right\} \]

\[ \lim_{n} \frac{R_{\pi^o}(n)}{\ln n} = M^{BK}(\theta), \quad \text{for all } \theta_i \in \Theta, \text{ and } i = 1, \ldots, N \]
\[ \Pi_i : \quad \mu_i = \mu(f_i) = \int_{\text{Sp}(f)} xf(x)dx, \quad \text{and} \quad \mu^* = \mu^*(\{f_i\}) = \max_i \mu(f_i) \]

where:
\[ f_i \in \mathcal{F} \] a known family of probability densities on \( \mathbb{R} \), with finite mean \( \mu(f) \)
\( \text{Sp}(f) \) is the support of \( f \)

Let \( \mu_i^* = \mu^* = \max_i \mu_i \) and \( \Delta_i = \mu^* - \mu_i \geq 0 \)
\[\Pi_i : \quad \mu_i = \mu(f_i) = \int_{Sp(f)} x f(x) dx, \quad \text{and} \quad \mu^* = \mu^*(\{f_i\}) = \max_i \mu(f_i)\]

where:

- \(f_i \in \mathcal{F}\) a known family of probability densities on \(\mathbb{R}\), with finite mean \(\mu(f)\)
- \(Sp(f)\) is the support of \(f\)
- Let \(\mu_i^* = \mu^* = \max_i \mu_i\) and \(\Delta_i = \mu^* - \mu_i \geq 0\)

\[M_{BK}(\{f_i\}) = \sum_{i : \mu_i \neq \mu^*} \frac{\Delta_i}{\inf_{g \in \mathcal{F}} \{I(f_i, g) : \mu(g) \geq \mu^*\}}\]

\[I(f, g) = E_f \left[ \ln \left( \frac{f(X)}{g(X)} \right) \right]\]
$\Pi_i : \mu_i = \mu(f_i) = \int_{S_p(f)} xf(x) \, dx$, and $\mu^* = \mu^*(\{f_i\}) = \max_i \mu(f_i)$

where:

$f_i \in \mathcal{F}$ a known family of probability densities on $\mathbb{R}$, with finite mean $\mu(f)$

$S_p(f)$ is the support of $f$

Let $\mu_i^* = \mu^* = \max_i \mu_i$ and $\Delta_i = \mu^* - \mu_i \geq 0$

$$M_{BK}(\{f_i\}) = \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\inf_{g \in \mathcal{F}} \{I(f_i, g) : \mu(g) \geq \mu^*\}}$$

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

$$u_{BK}^i(n, t, \hat{f}_t^i) = \sup_{g \in \mathcal{F}} \left\{ \mu(g) : I(\hat{f}_t^i, g) < \frac{\ln n}{t} \right\}$$
Non - Parametric Extention of Burnetas - Katehakis 1996

\[ \Pi_i : \mu_i = \mu(f_i) = \int_{Sp(f)} xf(x)dx, \text{ and } \mu^* = \mu^*(\{f_i\}) = \max_i \mu(f_i) \]

where:

\[ f_i \in \mathcal{F} \text{ a known family of probability densities on } \mathbb{R}, \text{ with finite mean } \mu(f) \]

\[ Sp(f) \text{ is the support of } f \]

Let \( \mu_i^* = \mu^* = \max_i \mu_i \) and \( \Delta_i = \mu^* - \mu_i \geq 0 \)

\[ M_{BK}(\{f_i\}) = \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\inf_{g \in \mathcal{F}} \{ I(f_i, g) : \mu(g) \geq \mu^* \}} \]

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

\[ u_{BK}^i(n, t, \hat{f}_t) = \sup_{g \in \mathcal{F}} \left\{ \mu(g) : I(\hat{f}_t, g) < \ln \frac{n}{t} \right\} \]

**UM** Sufficient conditions on the estimators \( \{\hat{f}_k^i\} \):

for all choices of \( \{f_i\} \subset \mathcal{F} \) and all \( \epsilon > 0, \delta > 0 \), the following hold for each \( i \), as \( k \to \infty \).

**C1**: \( \Pr \left( J(\hat{f}_k^i, \mu^* - \epsilon) < J(f_i, \mu^* - \epsilon) - \delta \right) = o(1/k) \).

**C2**: \( \Pr \left( u_{BK}^i(k, j, \hat{f}_j^i) \leq \mu_i - \epsilon, \text{ for some } j \in \{n_0, \ldots, k\} \right) = o(1/k) \).

\[ J(f, c) = \inf_{g \in \mathcal{F}} \{ I(f, g) : \mu(g) > c \} \]
$X^i_k$ are iid $N(\mu, \sigma^2)$ with

- unknown $\mu_i$ and **known** $\sigma_i^2$, we have $\Theta_i = \mathbb{R}$, $\hat{\mu}_i(n) = \overline{X}^i_{T^i_\pi(n)}$ and

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

- unknown both $\mu_i$ and $\sigma_i^2$, the indices below define a $\pi^0$ in $UM$

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

**C2 - open problem from 1996**

Auer et al (2002):

**UCB1:** (for arbitrary distributions $P_1, \ldots, P_m$ with support in $[0, 1]$)

$$u^A_i(\hat{\theta}_i(n), n) = \overline{X}^i_{T^i_\pi(n)} + (2 \log(n)/T_i(n))^{1/2}$$

**UCB1-Normal**

$$u^A_i(\hat{\theta}_i(n), n) = \overline{X}^i_{T^i_\pi(n)} + 4 \hat{\sigma}_i(T^i_\pi(n))(\log(n)/T^i_\pi(n))^{1/2}$$
$R_{\pi_A}(n) = M_{\pi_A}(\theta) \log n + o(\log n)$

$R_{\pi_B}(n) = M_{\pi_B}(\theta) \log n + o(\log n)$

$\Pi_1 : X_{11}, X_{12}, \ldots \text{ iid } N(8.1, 1)$
$\Pi_2 : X_{21}, X_{22}, \ldots \text{ iid } N(8.1, 4)$
$\Pi_3 : X_{31}, X_{32}, \ldots \text{ iid } N(7.9, 0.5)$
$\Pi_4 : X_{41}, X_{42}, \ldots \text{ iid } N(7, 3)$
$\Pi_5 : X_{51}, X_{52}, \ldots \text{ iid } N(-1, 1)$
$\Pi_6 : X_{61}, X_{62}, \ldots \text{ iid } N(0, 4)$
Regret Comparison

\[ \frac{R_{\pi BK}(n)}{R_{\pi Au}(n)} \]
Univariate discrete distribution:

\[ f_i(x, \theta_i) = \theta_i x \mathbb{1}\{X_i = r_{ix}\}, \quad x = 1, \ldots, d_i, \]

where \( \{r_{i1}, \ldots, r_{idi}\} \) are known constants, unknowns:

\( \underline{\theta}_i = (\theta_{i1}, \ldots, \theta_{idi}) \in \Theta_i = \{q \in \mathbb{R}^{di} : q_k \geq 0 \land \sum_{k=1}^{di} p_k = 1\} \)

Estimates: \( \hat{\theta}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{idi}) \)

\[
u^{BK}(\hat{\theta}_i) = \max_{q_i \geq 0} \left\{ \sum_{k=1}^{di} r_{ik} q_{ik} : \Pi(\hat{\theta}_i, q_i) \leq \frac{\ln n}{T^i_{\pi_0}(n)} \land \sum_{k=1}^{di} q_{ik} = 1 \right\}.
\]

\[
M^{BK}(\underline{\theta}) = \sum_{i=1}^{N} \Delta_i(\underline{\theta}_i) / \min_{q_i \geq 0} \{\Pi(\underline{\theta}_i, q_i) : \sum_{k=1}^{di} r_{ik} q_{ik} > \mu(\underline{\theta}_*^*) \land \sum_{k=1}^{di} q_{ik} = 1\}.
\]

For this problem Honda and Takemura (2011) constructed optimal policies, cyclic and randomized, that are simpler to implement.
Univariate discrete distribution:

\[ f_i(x, \theta_i) = \theta_i x, \mathbb{1}\{X_i = r_i x\}, \ x = 1, \ldots, d_i, \]

where \( \{r_{i1}, \ldots, r_{id_i}\} \) are known constants, unknowns:

\[ \theta_i = (\theta_{i1}, \ldots, \theta_{id_i}) \in \Theta_i = \{q \in \mathbb{R}^{d_i} : q_k \geq 0 \& \sum_{k=1}^{d_i} p_k = 1\} \]

Estimates: \( \hat{\theta}_i = (\hat{\theta}_{i1}, \ldots, \hat{\theta}_{id_i}) \)

\[
u^{BK}(\hat{\theta}_i) = \max_{q_i \geq 0} \left\{ \sum_{k=1}^{d_i} r_{ik} q_{ik} : \Pi(\hat{\theta}_i, q_i) \leq \frac{\ln n}{T^i_{\pi_0}(n)} \& \sum_{k=1}^{d_i} q_{ik} = 1 \right\} .
\]

\[
M^{BK}(\theta) = \sum_{i=1}^{N} \Delta_i(\theta_i) / \min_{q_i \geq 0} \{\Pi(\theta_i, q_i) : \sum_{k=1}^{d_i} r_{ik} q_{ik} > \mu(\theta^*) \& \sum_{k=1}^{d_i} q_{ik} = 1\}.
\]

For this problem Honda and Takemura (2011) constructed optimal policies, cyclic and randomized, that are simpler to implement.


\[ \tilde{R}_\pi(n) = n\mu^* - \sum_{i=1}^{N} \mu_i \left[ T^i_\pi(n) \right] = \sum_{i=1}^{N} \Delta_i T^i_\pi(n), \]

**Definition:** For a function \( g(n) \), a policy \( \pi \) is \( g \)-good if for every set of bandit distributions \( \mathcal{F} \), there exists a constant \( C_\pi(\mathcal{F}) < \infty \) such that

\[ \lim_{n} \frac{\tilde{R}_\pi(n)}{g(n)} \leq C_\pi(\mathcal{F}) \text{ (a.s) as } n \to \infty. \]  

A policy is \( g \)-good if

\[ \tilde{R}_\pi(n) = O(g(n)) \text{ (a.s), } n \to \infty \]

Trivially, policies exist that are \( n \)-good (i.e., \( \tilde{R}_\pi(n) = O(n) \text{ (a.s.)} \)), for example any policy that samples all populations at constant rate \( 1/N \).

**Theorem 1.** For \( g \), an unbounded, positive, increasing, concave, differentiable, sub-linear function, there exist \( g \)-good policies.  

The proof of this theorem is given by example when we demonstrate two \( g \)-good policies: the \( g \)-Forcing and the \( g \)-ISM index policies.
**g-Forcing policy:** A policy \( \pi^F_g \) that first samples each bandit once, then for \( t \geq N \),

\[
\pi^F_g(t + 1) = \begin{cases} 
\arg \max_i \overline{X}^i_{T^i_{\pi^F_g}(t)} & \text{if } \min_i T^i_{\pi^F_g}(t) \geq g(t), \\
\arg \min_i T^i_{\pi^F_g}(t) & \text{else.}
\end{cases}
\]  

(3)

**Theorem** For a policy \( \pi^F_g \) as above \( \pi^F_g \) is \( g \)-good, and

\[
P \left( \lim_{n} \frac{\tilde{R}^F_{\pi^F_g}(n)}{g(n)} = S_\Delta \right) = 1.
\]

Where

\[
S_\Delta = \sum_{i: \mu_i \neq \mu^*} \Delta_i.
\]

The value \( S_\Delta \) in some sense represents the pseudo-regret incurred each time the sub-optimal bandits are all activated once. The next result states that g-Forcing policies satisfy the conditions of Theorem 1.
This theorem actually follows immediately from the following, much stronger statement:

For a policy $\pi^F_g$ the following is true: For every $\epsilon > 0$, almost surely there exists a $N_\epsilon < \infty$ such that, for all $n \geq N_\epsilon$,

$$g(n)S_\Delta - \epsilon \leq \tilde{R}_{\pi^F_g}(n) \leq [g(n)]S_\Delta.$$  

$$S_\Delta = \sum_{i: \mu_i \neq \mu^*} \Delta_i.$$
This theorem actually follows immediately from the following, much stronger statement:

For a policy $\pi^F_g$ the following is true: For every $\epsilon > 0$, almost surely there exists a $N_\epsilon < \infty$ such that, for all $n \geq N_\epsilon$,

$$g(n)S_\Delta - \epsilon \leq \tilde{R}_{\pi^F_g}(n) \leq \lfloor g(n) \rfloor S_\Delta.$$

$S_\Delta = \sum_{i: \mu_i \neq \mu^*} \Delta_i$.

In fact, for most large $n$:

$$\tilde{R}_{\pi^F_g}(n) = \lfloor g(n) \rfloor S_\Delta$$
**Class of g-Index Policies**

**g-ISM index policy:** A policy \( \pi_g \) that first samples each bandit once, then for \( t \geq N \),

\[
\pi_g(t + 1) = \arg\max_i u_i(t, T_{\pi_g}^i(t)) = \arg\max_i \left( \frac{\bar{X}_{T_{\pi_g}^i}(t)}{T_{\pi_g}^i(t)} + \frac{g(t)}{T_{\pi_g}^i(t)} \right).
\]

**Theorem** For a policy \( \pi_g \) as above, if \( N^* = 1 \),

\[
\mathbb{P} \left( \lim_{n} \frac{\tilde{R}_{\pi,g}(n)}{g(n)} = N - 1 \right) = 1.
\]

**FINITE HORIZON Bounds**

For each sub-optimal \( i, \forall \epsilon \in (0, \Delta_i/2), \exists \) (a.s.) a finite constant \( C^i_\epsilon \) such that for \( n \geq N' \),

\[
T_{\pi_g}^i(n) \leq \frac{g(n)}{\Delta_i - 2\epsilon} + C^i_\epsilon.
\]

If \( N^* = 1 \), for each sub-optimal \( i \neq i^* \), \( \forall \epsilon \in (0, \min_j \neq i^* \Delta_j/2) \), \( \exists \) (a.s.) some finite \( N' \) such that for \( n \geq N' \),

\[
\frac{g(n)}{(1 + \epsilon)(\Delta_i + 2\epsilon) + 2\epsilon} \leq T_{\pi_g}^i(n).
\]
• Unknown Variance: $\pi^*$ an index policy based on $u_i^{CHK}(n)$

$$u_i^{CHK}(n) = \overline{X}_{T_i(n)} + \hat{\sigma}_{T_i(n)}^i \sqrt{n \frac{2}{T_i(n) - 2}} - 1$$

$$u_i^{BK}(n) = \overline{X}_{T_i(n)} + \hat{\sigma}_{T_i(n)}^i \sqrt{n \frac{2}{T_i(n)}} - 1$$

Results:
• Unknown Variance: $\pi^*$ an index policy based on $u^\text{CHK}_i(n)$

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

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

Results:

$$\lim_{n} \frac{R_{\pi^{\text{CHK}}}(n)}{\ln n} = M^{\text{BK}}(\mu, \sigma^2)$$
• Unknown Variance: $\pi^*$ an index policy based on $u_{i}^{CHK}(n)$

$$u_{i}^{CHK}(n) = \overline{X}_{T^i(n)} + \hat{\sigma}_{T^i(n)}^i \sqrt{n\overline{T^i(n^2)} - 1}$$

$$u_{i}^{BK}(n) = \overline{X}_{T^i(n)} + \hat{\sigma}_{T^i(n)}^i \sqrt{n\overline{T^i(n^2)} - 1}$$

Results:

$$\lim_{n} \frac{R_{\pi^{CHK}}(n)}{\ln n} = M^{BK}(\mu, \sigma^2)$$

$$R_{\pi^{CHK}}(n) \leq \sum_{i: \mu_i \neq \mu^*} \left( \frac{2 \ln n}{\ln \left(1 + \frac{\Delta^2_i}{\sigma^2_i} \cdot \frac{(1-\epsilon)^2}{(1+\epsilon)} \right)} + \sqrt{\frac{\pi}{2\epsilon}} \cdot \frac{8\sigma^3_i}{\Delta^3_i \epsilon^3} \ln \ln n + \frac{8}{\epsilon^2} + \frac{8\sigma^2_i}{\Delta^2_i \epsilon^2} + 4 \right) \Delta_i.$$
Numerical Regret Comparison: Figure 1 shows the results of a small simulation study, implementing policies $\pi_{\text{CHK}}, \pi_{\text{ACF}},$ and $\pi_{\text{G}}$ a ‘greedy’ policy that always activates the bandit with the current highest average. The simulation was done on a set of six populations, with means and variances given in the table below.

<table>
<thead>
<tr>
<th>$\mu_i$</th>
<th>8</th>
<th>8</th>
<th>7.9</th>
<th>7</th>
<th>-1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_i^2$</td>
<td>1</td>
<td>1.4</td>
<td>0.5</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

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

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

Figure: Left: Plots of $B_{\pi_{ACF}}(n)$ and $B_{\pi_{CHK}}(n)$. Right: Convergence of $R_{\pi_{CHK}}(n)/\ln(n)$ to $M^{BK}(\mu, \sigma^2)$.
For each $i$, $f_i \in \mathcal{F}$ Uniform on $[a_i, b_i]$, $\mu(f_i) = (a_i + b_i)/2$.

$$\hat{a}_t^i = \min_{t' \leq t} X_t^{i'} \quad \& \quad \hat{b}_t^i = \max_{t' \leq t} X_t^{i'}$$

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

$$u_{\text{BK}}^i(n, t, \hat{f}_t^i) = \hat{a}_t^i + \frac{1}{2} \left( \hat{b}_t^i - \hat{a}_t^i \right) n^{1/t}.$$
Uniform Bandits

For each $i$, $f_i \in \mathcal{F}$ Uniform on $[a_i, b_i]$, $\mu(f_i) = (a_i + b_i)/2$.

$$\hat{a}_i^t = \min_{t' \leq t} X_{t'}^i \quad \& \quad \hat{b}_i^t = \max_{t' \leq t} X_{t'}^i$$

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

$$u^{BK}_i (n, t, \hat{f}_t^i) = \hat{a}_i^t + \frac{1}{2} \left( \hat{b}_i^t - \hat{a}_i^t \right) n^{1/t}.$$ 

$$u^{CK}_i (n, t, \hat{f}_t^i) = \hat{a}_i^t + \frac{1}{2} \left( \hat{b}_i^t - \hat{a}_i^t \right) n^{\frac{1}{t-2}}$$
References to Recent Work

arXiv:1505.02865 [pdf, ps, other]
Asymptotic Behavior of Minimal–Exploration Allocation Policies: Almost Sure, Arbitrarily Slow Growing Regret
Wesley Cowan, Michael N. Katehakis
Subjects: Machine Learning (stat.ML); Learning (cs.LG)

arXiv:1505.01918 [pdf, ps, other]
An Asymptotically Optimal Policy for Uniform Bandits of Unknown Support
Wesley Cowan, Michael N. Katehakis
Subjects: Machine Learning (stat.ML); Learning (cs.LG)

arXiv:1504.05823 [pdf, other]
Normal Bandits of Unknown Means and Variances: Asymptotic Optimality, Finite Horizon Regret Bounds, and a Solution to an Open Problem
Wesley Cowan, Junya Honda, Michael N. Katehakis
Comments: 15 pages 3 figures
Subjects: Machine Learning (stat.ML); Learning (cs.LG)
References


References


