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 $\pi^*$ (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’:
  \[ \underline{f} = \{f_1, f_2, \ldots, f_N\} \subset \mathcal{F} \]
- May sample i.i.d. from any bandit: $X_1^i, X_2^i, \ldots \sim f_i$
- Given $t$ samples of $i$, may construct estimator $\hat{f}_t^i$
- Sequential sampling policy $\pi$, $\pi(n) = i$ samples $i$ at time $n$
  - $T^i_\pi(n)$: # of samples of $i$ at global time $n$ (global $n$ vs local $T^i_\pi(n)$)
- Score functional $s : \mathcal{F} \mapsto \mathbb{R}$.
- Optimal bandits: $s(f_i^*) = s^* = \max_j s(f_j)$.

General Goal:
- A policy $\pi$ that samples optimal bandits as often as possible
- Efficiently balance exploration vs exploitation
What is Good?

Let \( O(f) = \{ i : s(f_i) = s^*(f) \} \), \( B(f) = \{ i : s(f_i) < s^*(f) \} \) be the set of optimal, sub-optimal bandits.

Basic Principle: Activations of optimal bandits cannot be regretted.

**Definition (Uniformly Fast Policies)**

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

\[
\sum_{i \in B(f)} \mathbb{E}_f \left[ T^i_\pi(n) \right] = o(n^{\alpha}),
\]

- Regret:

\[
R_\pi(n) = R_\pi(n; f) \sum_{i \in B(f)} (s^*(f) - s(f_i)) \mathbb{E}_f \left[ T^i_\pi(n) \right]
\]

Structure of Bandit Space

• KL-Divergence as ‘distance/similarity’ in $\mathcal{F}$:

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

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

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

• $K_f(\rho)$: Distance to nearest $\rho$-better $g$
Assume the following conditions hold, for any \( f \in \mathcal{F} \), and all \( \epsilon, \delta > 0 \).

\begin{itemize}
  \item [\diamondsuit \textbf{Condition B1:}] \( \forall f \in \mathcal{F}, \rho \in s(\mathcal{F}), \exists \tilde{f} \in \mathcal{F} : s(\tilde{f}) > \rho \) and \( I(f, \tilde{f}) < \infty \).
  \item [\diamondsuit \textbf{Condition B2:}] \( s \) is continuous at each \( f \in \mathcal{F} \), with respect to \( I \).
\end{itemize}

\textbf{Theorem (Lower Bound on Sub-Optimal Activations)}

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

\[
\liminf_{n} \frac{\mathbb{E}_{f}[T_{\pi}^{i}(n)]}{\ln n} \geq \frac{1}{K_{f_{i}}(s^{*})}.
\]

Are there policies (‘asymptotically optimal’) that achieve this lower bound?
Realizing the Bound

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

$$\lim_{n} \mathbb{E}[T_{\pi}^{i}(n)]/ \ln n = 1/K_{f_{i}}(s^{*})$$

Let $\nu$ 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**: $K_{f}(\rho)$ is continuous w.r.t $\rho$, and w.r.t $f$ under $\nu$.
- **Condition R2**: $P_{f}(\nu(\hat{f}_{t}^{f}, f) > \delta) \leq o(1/t)$.
- **Condition R3**: For some sequence $d_{t} = o(t)$ (independent of $\epsilon, \delta, f$),

$$P_{f}(\delta < K_{\hat{f}_{t}^{f}}(s(f) - \epsilon)) \leq e^{-\Omega(t)} e^{-(t-d_{t})\delta},$$

where the dependence on $\epsilon$ 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:** $K_f(\rho)$ is continuous w.r.t $\rho$, and w.r.t $f$ under $\nu$. It characterizes, in some sense, the structure of $\mathcal{F}$ as smooth. To the extent that $K_f(\rho)$ 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 $\epsilon, \delta, f$),

\[
\mathbb{P}_f(\delta < K_{\hat{f}_t}(s(f) - \epsilon)) \leq e^{-\Omega(t)}e^{-(t-d_t)\delta},
\]

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 $\delta$, can be relaxed somewhat, but such a bound frequently seems to exist in practice, for natural choices of $\hat{f}_t$.***
Policy $\text{UCB-}(\mathcal{F}, s, \hat{f}_t, \tilde{d})$

- 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:

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

Policy $\pi^*$ (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

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

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)
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 $\inf_{g \in \mathcal{F}} \{K_g(s^* - \epsilon) : \nu(g, f_i) \leq \delta\} > 0$:

$$
\mathbb{E}[T^i_{\pi^*}(n)] \leq \frac{\ln n}{\inf_{g \in \mathcal{F}} \{K_g(s^* - \epsilon) : \nu(g, f_i) \leq \delta\}} + o(\ln n)
$$

$$
+ \sum_{t=n_0N}^n \mathbb{P} \left( \nu(\hat{f}_t^i, f_i) > \delta \right)
$$

$$
+ \sum_{t=n_0N}^n \sum_{k=n_0}^t \mathbb{P} \left( u_{i^*}(t, k) \leq s^* - \epsilon \right).
$$
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 \(\Delta\), for all \(t\), then

\[
\lim_{n \to \infty} \frac{\mathbb{E} \left[ T_{\pi^*}(n) \right]}{\ln n} = \frac{1}{K_{f_i}(s^*)}, \quad \forall f \in \mathcal{F}, \text{ and } \forall i \text{ suboptimal.}
\]
Applications: Separable Pareto Models

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

\( X \sim \text{Pareto}(\alpha, \beta) \), \( X \) is distributed over \([\beta, \infty)\),
with \( \mathbb{E}[X] = \frac{\alpha \beta}{(\alpha - 1)} \) 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 \( \beta \), and a decreasing function of \( \alpha \).

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 \( \alpha \) for \( \alpha > \ell \), and \( b \) to be a positive, continuous, non-decreasing function of \( \beta^* \).

* 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 \( \alpha \) and increasing \( \beta \).
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

\[
s(\alpha, \beta) = \frac{\alpha \beta}{\alpha - 1},
\]

with \( a(\alpha) = \frac{\alpha}{\alpha - 1} \) and \( b(\beta) = \beta \).

ii) For unrestricted Pareto distributions, the score function

\[
s(\alpha, \beta) = \frac{1}{\alpha},
\]

leads to the controller’s goal to be to find the bandit with minimal \( \alpha \). In this case, \( a(\alpha) = \frac{1}{\alpha} \) 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.
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

\[
\hat{\beta}_t = \min_{n=1,...,t} X_k, \\
\hat{\alpha}_t = \frac{t - 1}{\sum_{k=1}^{t} \ln \left( \frac{X_k}{\hat{\beta}_t} \right)}. \tag{1}
\]

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 \( \delta \), and \( L^\pm (\delta) \to 1 \) as \( \delta \to 0 \).
Policy $\pi_{p,s}^*$ (UCB-PARETO)

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

ii) for $n \geq 3N$, sample from bandit $\pi_{p,s}^*(n + 1) = \arg\max_i u_i \left( n, T_{\pi_{p,s}^*}^i(n) \right)$ breaking ties uniformly at random, where

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

**Theorem**

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

$$\lim_{n \to \infty} \frac{\mathbb{E} \left[ T_{\pi_{p,s}^*}^i(n) \right]}{\ln n} = \frac{1}{\frac{1}{\alpha_i} a^{-1} \left( \frac{s^*}{b(\beta_i)} \right) - \ln \left( \frac{1}{\alpha_i} a^{-1} \left( \frac{s^*}{b(\beta_i)} \right) \right) - 1}.$$
Applications: General Uniform Models

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

- 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:

\[ s_\mu(a, b) = (a + b)/2. \]
\[ F = \left\{ f_{a,b}(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b : -\infty < a < b < \infty \right\} \]

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

\[
\hat{a}_t = \min_{n=1,\ldots,t} X_n \quad \hat{b}_t = \max_{n=1,\ldots,t} X_n.
\]

- Index of Optimal Policy \( \pi^*, \tilde{d} = 2 \):

\[
u_i(n,t) = s(\hat{a}_i, \hat{a}_i + n^{t-2} (\hat{b}_i - \hat{a}_i)).
\]

with the particular case for \( s_\mu \):

\[
u_i(n,t) = \hat{a}_i + \frac{1}{2} n^{t-2} (\hat{b}_i - \hat{a}_i).
\]
Policy $\pi^*$ is asymptotically optimal, and for all $\{f_i = f_{a_i,b_i}\} \subset \mathcal{F}$, for all sub-optimal $i$:

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

with the particular case for $s_\mu$:

$$\lim_{n} \frac{\mathbb{E} \left[ T_{\pi^*}^i (n) \right]}{\ln n} = \frac{1}{\ln \left( \frac{2s^* - 2a_i}{b_i - a_i} \right)}.$$
Applications: Normal, Unknown $\mu_i, \sigma_i$, 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: $s(f_{\mu, \sigma}) = \mathbb{E}_f [X] = \mu$.
- Standard Estimators: $\hat{\mu}_t$ and $\hat{\sigma}_t^2$.
- Index of Optimal Policy $\pi^* = \pi_{\text{CHK}}$, $\tilde{d} = 2$:

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

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

$$\lim_{n} \frac{\mathbb{E} \left[ T_{\pi_{\text{CHK}}}^i (n) \right]}{\ln n} = \frac{2}{\ln \left( 1 + \frac{(\mu^* - \mu_i)^2}{\sigma_i^2} \right)}.$$

(Cowan, Honda, and Katehakis 2015)
Normal, Minimize Variance, known $\mu_i$,

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

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

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

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 $i$:

$$\lim_{n \to \infty} \frac{\mathbb{E} [T^i_{\pi^*}(n)]}{\ln n} = \frac{2}{\frac{\sigma_i^2}{\sigma_*^2} - \ln \left( \frac{\sigma_i^2}{\sigma_*^2} \right) - 1}.$$
Normal \( \kappa \)-Threshold Probability, known \( \sigma_i \)

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

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

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

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

\[
\lim_{n \to \infty} \frac{\mathbb{E} \left[ T_{\pi^*}^i(n) \right]}{\ln n} = \frac{2}{\left( \frac{\kappa - \mu_i}{\sigma_i} - \Phi^{-1} (1 - p^*) \right)^2}.
\]
Final Comments: Past and Current Work

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

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

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

Burnetas - Katehakis 1996: $f(x; \theta_i)$ unknown multi-dim (vector) $\theta_i \in \Theta$

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

$$\mathbb{K}_i^{BK}(\theta^*) = \inf_{\theta'_i \in \Theta_i} \{I(\theta_i, \theta'_i) : \mu(\theta'_i) > \mu(\theta^*)\}$$

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

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

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

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

$$\liminf_n \frac{\mathbb{E}[T^n_i(n)]}{\ln n} \geq \frac{1}{\mathbb{K}_i(s^*)}, \forall UF \pi$$
LR-UM Policies $\phi^*$ 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 $\Pi_n$ (initial sampling)
- sample mean estimates: $\hat{\mu}_i(n) = \mu(\hat{\theta}_i)$
- First UCBs: $g_i^n(\hat{\theta}_i)(= u_i^{LR}(\hat{\theta}_i)) = \inf\lambda\{\lambda > \mu(\hat{\theta}_i) : \mathbb{I}(\hat{\theta}_i, \lambda) \geq a_{ni}\}$
- 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

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

$$\lim_{n} \frac{\mathbb{E} \left[ T^i_{\pi^{LR}(n)} \right]}{\ln n} = \frac{1}{K_i(\mu^*)}, \quad \forall \text{ non-optimal } i$$
BK-UM Policies $\pi^*$ 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 $T_{\pi^*}^i(n) > 0$ for all $i$. Initial estimates $\hat{\theta}_i(n) = \hat{\theta}_i(T_{\pi^*}^i(n))$

- 2-nd UCBs:

$$u_{BK}^i(n, t) = \sup_{\theta' \in \Theta_i} \left\{ \mu(\theta'_{i}) : \mathbb{I}(\hat{\theta}_i, \theta'_{i}) < \frac{\ln n}{t} \right\}$$

- 2-nd UCB index based Efficient Policies:

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

breaking ties uniformly at random

- Then

$$\lim_{n} \frac{\mathbb{E} \left[ T_{\pi^{BK}}^i(n) \right]}{\ln n} = \frac{1}{K_i(\mu(\theta^*))}, \forall \ non-optimal \ i$$

$\pi^*$ is a pure index policy
Asymptotically Optimal Policies of Cowan and Katehakis 2015

Policy UCB-\( (F, s, \hat{f}_t, \tilde{d}, \nu) \) \( \pi^* \) 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 \( \hat{f}_i^t \) 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:

\[
\begin{align*}
  u_i(n, t) &= \sup_{g \in F} \left\{ s(g) : I(\hat{f}_i^t, g) \leq \frac{\ln n}{t - \tilde{d}} \right\} \\
\end{align*}
\]

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

\[
\pi^*(n + 1) = \arg \max_i \left\{ u_{i}^{CK}(n, T_{\pi^*}^i(n)) \right\} 
\]
breaking ties uniformly at random

- Then

\[
\lim_{n} \frac{\mathbb{E} \left[ T_{\pi^*}^i(n) \right]}{\ln n} = \frac{1}{K_i(s^*)}, \quad \forall \text{ non-optimal } i
\]

\( \pi^* \) is a pure index policy
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

$$g_{nk}^i = g_{nk}^i(\hat{\mu}_i(k), a_{nk}) = \hat{\mu}_i(k) + \sigma(2a_{nk})^{1/2}$$
Asympt. Optimal UCB Policies for Normal Populations: \( X^i_k \) 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

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

Estimates \( \hat{\mu}_i(k) = \hat{\theta}_i(k) = \sum_{m=1}^k X^i_m/k \) define

\[
g^i_{nk} = g^i_{nk}(\hat{\mu}_i(k), a_{nk}) = \hat{\mu}_i(k) + \sigma(2a_{nk})^{1/2}
\]

For \( n + 1 > N \) compute: \( j: n + 1 = mN + j \) and \( j^*_n: \)

\[
\hat{\mu}_{j^*_n} = \max\{\hat{\mu}_i(n) : T^i_{\pi_{LR}}(n) > \delta n\}
\]

\[
\pi_{LR}(n + 1) = \begin{cases} 
  j & \text{if } \hat{\mu}_{j^*_n} < g^i_{nk}(\hat{\mu}_i(k), a_{nk}) \\
  j^*_n & \text{otherwise}
\end{cases}
\]
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 $\hat{\mu}_i(k) = \hat{\theta}_i(k) = \sum_{m=1}^{k} X_i^m/k$ 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.
Asympt. Optimal UCB Policies for Normal Populations: \( X^i_k \) are iid \( N(\mu_i, \sigma^2_i) \)

**Katehakis and Robbins (1995):** \( \mu_i \) unknown and \( \sigma^2_i \) known

A policy \( \pi_g \) that first samples each bandit once, then for \( n \geq N + 1 \),

\[
u^K_{iR}(\hat{\mu}_i(n), n) = \bar{X}^i_{T\pi}(n) + \sigma_i(2 \log n/T^i(n))^{1/2}
\]

**Burnetas and Katehakis (1996):** both \( \mu_i \) and \( \sigma^2_i \) unknown

A policy \( \pi_g \) that first samples each bandit once, then for \( t \geq N + 1 \),

\[
u^B_{iK}(\hat{\theta}_i(n), n) = \bar{X}^i_{T\pi}(n) + \hat{\sigma}_i(T^i(n))(n^2/(T^i(n)) - 1)^{1/2}
\]

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

**Cowan, Honda and Katehakis (2015):** both \( \mu_i \) and \( \sigma^2_i \) unknown

A policy \( \pi_g \) that first samples each bandit once, then for \( t \geq N + 1 \),

\[
u^C_{iHK}(\hat{\theta}_i(n), n) = \bar{X}^i_{T\pi}(n) + \hat{\sigma}_i(T^i(n))(n^2/(T^i(n) - 2) - 1)^{1/2}
\]
Existence of UF policy $\pi_{ACF}$:

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

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

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

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

And the bound

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

with

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

\[
C_{ACF}(\mu) = \left(1 + \frac{\pi^2}{2}\right) \sum_{i=1}^N \Delta_i.
\]
Existence of UF policy $\pi_{ACF}$:

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

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

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

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

And the bound

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

with

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

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

$$
R_{\pi_{ACF}}(n) \leq M_{ACF}(\mu, \sigma^2) \ln n + o(\ln n). \ \ln n = o(n^\alpha) \text{ for all } \alpha > 0 \text{ and } R_{\pi_{ACF}}(n) \geq 0,
$$

i.e., $\pi_{ACF}$ is uniformly fast convergent.
Regret Comparison

\[ R_{\pi Au}(n) = M_{Au}(\theta) \log n + o(\log n) \]

\[ R_{\pi BK}(n) \geq O(\sqrt{n}) \]

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

Regret Comparison - Continued

\[ \frac{R_{\pi BK}(n)}{R_{\pi Au}(n)} \]
Recent Results

- **Unknown Variance:** $\pi^*$ an index policy based on $u_{i}^{CHK}(n)$

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


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

Burnetas and Katehakis (1996)

**Results:**

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

$$R_{\pi_{CHK}}(n) \leq \sum_{i: \mu_i \neq \mu^*} \left( \frac{2 \ln n}{\ln \left( 1 + \frac{\Delta_i^2 (1-\epsilon)^2}{\sigma_i^2 (1+\epsilon)} \right)} + \sqrt{\frac{\pi}{2e} \frac{8\sigma_i^3}{\Delta_i^3 \epsilon^3}} \ln \ln n + \frac{8}{\epsilon^2} + \frac{8\sigma_i^2}{\Delta_i^2 \epsilon^2} + 4 \right) \Delta_i.$$
Figures 1 & 2 show the results of a small simulation study, implementing policies $\pi_{\text{CHK}}, \pi_{\text{ACF}},$ and $\pi_G$ 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.

<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 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.

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

$$M^{BK}(\underline{\theta}) = \sum_{i \in \mathcal{B}(\underline{\theta})} 1/ \inf_{\theta'_i \in \Theta_i} \{I(\theta_i, \theta'_i) : \mu(\theta'_i) > \mu(\theta^*)\}$$
Figure 2 shows first (left) a comparison of the theoretical bounds on the regret, \( B_{\pi_{ACF}}(n) \) and \( B_{\pi_{CHK}}(n) \) representing their theoretical regret bounds respectively, for the means and variances indicated in the table below. Additionally, Figure 2 (right) shows the convergence of \( R_{\pi_{CHK}}(n) / \ln(n) \) to the theoretical lower bound \( M^{BK}(\mu, \sigma^2) \).

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.
Thompson Sampling for Normal Populations - unknown $\mu_i$ & $\sigma^2_i$

Honda and Takemura 2013

Policy $\pi_{TS}$ (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 $\mu_i$, given $\left(\bar{X}_{T^i_{\pi}(n)}, \hat{\sigma}^2_i(T^i_{\pi}(n))\right)$, and a prior for $(\mu_i, \sigma^2_i) \propto (\sigma^2_i)^{-1-\alpha}$.

iii) Then, take

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

$$\lim_{n} \frac{R_{\pi_{TS}}(n)}{\ln n} = M^{BK}(\mu, \sigma^2), \quad \forall(\mu, \sigma^2)$$
Thompson Sampling - Normal with unknown $\mu_i$ & $\sigma^2_i$

Cowan, Honda and Katehakis 2015

Numerical Regret Comparison of $\pi_{CHK}$ and $\pi_{TS}$

$R_{\pi_{CHK}}(n)$ and $R_{\pi_{TS}}(n)$ for the parameters, of Table 1, left and Table 2, right.
For each $i$, $f_i \in F$ Uniform on $[a_i, b_i]$

$$
\mu(f_i) = \frac{a_i + b_i}{2}
$$

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

$$
u_{CK}^i(n, t, \hat{f}_t^i) = \hat{a}_t^i + \frac{1}{2} \left( \hat{b}_t^i - \hat{a}_t^i \right) n^{\frac{1}{t-2}} \text{ asymptotically optimal!}
$$

$$
\pi_{CK}(n + 1) = \arg \max_i \ u_{CK}^i(n, t, \hat{f}_t^i)
$$

$$
M_{BK}^{BK}([\{a_i, b_i\}]) = \sum_{i: \mu_i \neq \mu^*} \frac{\Delta_i}{\ln \left( 1 + \frac{2\Delta_i}{b_i - a_i} \right)}
$$
For each $i$, $f_i \in \mathcal{F}$ Uniform on $[a_i, b_i]$

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

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

$$
\pi_{CK}(n + 1) = \arg \max_i u_{CK}^i(n, t, \hat{f}_t^i)
$$

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

$$
u_{BK}^i(n, t, \hat{f}_t^i) = \hat{a}_t^i + \frac{1}{2} \left(\hat{b}_t^i - \hat{a}_t^i\right) n^{1/t} \quad \text{asymptotically optimal?}
$$
Short Time Horizon: Numerical regret comparison of $\pi_{CK}$, $\pi_{KR}$, and $\pi_{CHK}$, for the 6 bandits with parameters given in Table 1. Average values over 20,000 repetitions.

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_i$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>9.5</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
Longer Time Horizon: Numerical regret comparison of $\pi_{CK}$, $\pi_{KR}$, and $\pi_{CHK}$, for the 6 bandits with parameters given in Table 1.

Average values over 10,000 repetitions.

<table>
<thead>
<tr>
<th>$\pi_{KR}$</th>
<th>$u_{KR}^i(n, t) = \bar{X}_t^i + \hat{\sigma}^i(t)\sqrt{\frac{2 \ln n}{t}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_{CHK}$</td>
<td>$u_{CHK}^i(n, t) = \bar{X}_t^i + \hat{\sigma}^i(t)\sqrt{\frac{2}{n \frac{t}{t-2} - 1}}$</td>
</tr>
<tr>
<td>$\pi_{CK}$</td>
<td>$u_{CK}^i(n, t, \hat{j}_t^i) = \hat{a}_t^i + \frac{1}{2} (\hat{b}_t^i - \hat{a}_t^i) \sqrt{\frac{1}{n \frac{t}{t-2}}}$</td>
</tr>
</tbody>
</table>
References