Optimal Data Driven Policies under Constrained Multi-armed Bandit Observations

Odysseas Kanavetas
Mathematical Institute
Leiden University

Athens University of Economics and Business
Department of Statistics
December 2020

Beginnings: Robbins 1952

Example

Π_1 Π_2


$X_{11}, X_{12}, X_{13}, \ldots$ $X_{21}, X_{22}, X_{23}, \ldots$

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

Traditional Allocation

Assign 100 to Π_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 Π_2 observe 1,1,-1,1,1, , . . . ,1 with $\sum_1^{100} = 99 - 1 \rightarrow \hat{p}_2 = \frac{99}{100}$

``Killed" 90 on Π_1 to learn

CAN WE DO BETTER ?

Given 2 populations (treatments) that generate outcomes:

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

$$S_\pi(n) = \sum_{i=1}^{2} \sum_{k=1}^{T_\pi^i(n)} X_k^i \text{ (max)}$$

Defined measures of regret:

$$R_{\pi(n)} = n\mu^* - S_\pi(n) = n\mu^* - \sum_{i=1}^{2} \sum_{k=1}^{T_\pi^i(n)} X_k^i,$$

$\mu^* = \max\{\mu_1, \mu_2\}$

Early Work — Robbins 1952

Constructed a modified 'play the winner'(greedy)
(outside two sparse sequences of forced choices) policy, π_R:

$$S_{\pi_R}(n)/n \to \mu^* \text{ as } n \to \infty, \text{with probability one}$$

i.e.,

$$R_{\pi_R}(n) = o(n) \ (a.s.), \text{ as } n \to \infty.$$

Let $a_n^i > 0$ $(n = 1, 2, \ldots, i = 1, 2)$ be two sequences constants such that:

- for every fixed i, a_n^i is increasing in $n \geq 1$
- $$\lim_{n\to\infty} (\sum_{k=1}^{n} \mathbf{1}_{a_k^1 > 0} + \mathbf{1}_{a_k^2 > 0})/n = 0$$

For $n + 1 > 2$ let: j_n^*: $\hat{\mu}_{j_n^*} = \max\{\hat{\mu}_i(n)\ \}$

$$\pi^{LR}(n+1) = \begin{cases} 1 & \text{if } \exists k \ : \ a_k^1 = n+1 \\ 2 & \text{if } \exists k \ : \ a_k^2 = n+1 \\ j_n^* & \text{otherwise} \end{cases}$$

The MAB Problem: Lai - Robbins 1985

Π_i : $X_1^i, X_2^i, \ldots$ $f(x;\theta_i)$ unknown $\theta_i \in \Theta$.

where $f(.;.)$ is known univariate density w.r.t. some measure ν
Let $\underline{\theta} = (\theta_1, \ldots, \theta_N)$
$\mu_i = \mu(\theta_i) = \mathbb{E}X_1^i$, $\mu^* = \mu(\theta_{i*})$, $\Delta_i(\theta_i) = \mu^* - \mu(\theta_i)$,
$\mathbb{I}(\theta, \theta') = \int_{-\infty}^{\infty} \ln \frac{f(x;\theta)}{f(x;\theta')} f(x;\theta)\, \mathrm{d}\nu(x)$ be K-L divergence between $f(x;\theta)$ and $f(x;\theta')$.
Regret:

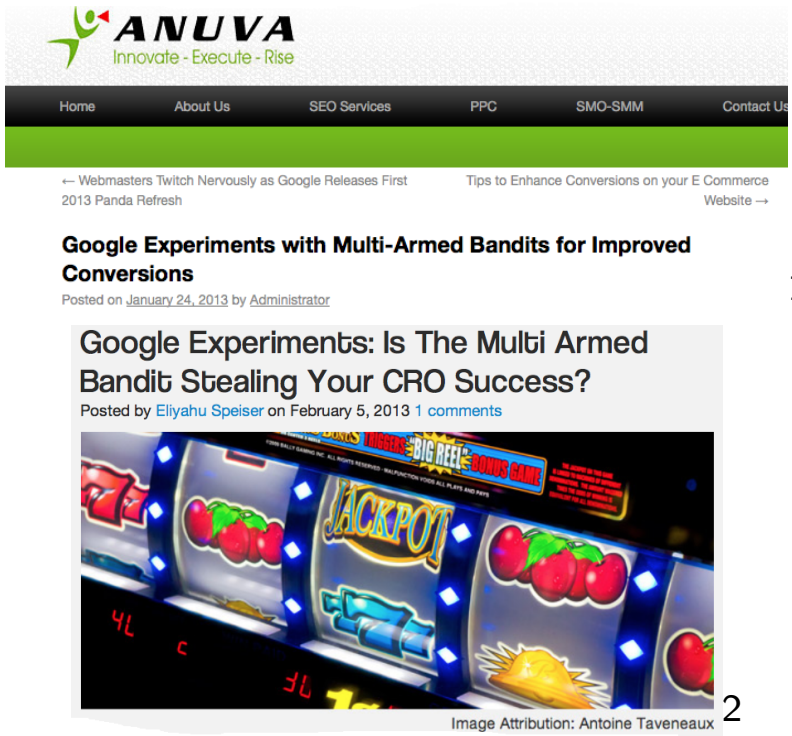
$$R_\pi(n) = R_\pi(n|\underline{\theta}) = n\mu^* - \mathbb{E}S_\pi(n) = \sum_{i=1}^{N} \Delta_i \mathbb{E}\left[T_\pi^i(n)\right]$$

Types of Policies

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

Google Website Optimizer - CRO 2013

Not only Multivariate Testing - Multi Armed Bandit (Google) Experiments



[1]
[2]

Conversion Rate Optimization (CRO)

[1] http://www.anuvatech.com/blog/google-experiments-with-multi-armed-bandits-for-improved-conversions/

[2] http://3doordigital.com/google-experiments-cro-success/

Motivation

- Medical treatments limited by
  - cost of materials
  - limited resources
- Different advertisements for a product
  - cost of media
  - each medium provides a limited capacity

Problem Setting

- Multi-armed bandit problem in the frequentist framework
- Average reward maximization (asymptotic)
- Sampling costs-constraints
- Unknown parameters of reward distributions

Definition of Multi-Constrained MAB Problem

- Sequential Sampling from k Statistical Populations
- Each period population i is sampled from:
  - X_i=random reward $\sim f_i(\cdot|\theta_i)$ (p.d.f. or p.m.f)
  - $\mu_i = \mathsf{E}(X_i|\theta_i)$
  - $\underline{\theta} = (\theta_1, \ldots, \theta_k)$
  - L types of resources
  - c_i^j= type-j units of used resources per sampling j=1,...,L
  - C_0^j=type-j units of available resources (on average basis)
- Objective: Identify a sampling policy for which
  - The long run expected average used recources per period of each type does not exceed the corresponding type of C_0.
  - The long run expected average reward per sample is maximized.

A Linear Programming Formulation

x_j=Fraction of time periods allocated to sampling from population j

Primal

$$
\begin{aligned}
z(\underline{\theta}) \quad &= \max \textstyle\sum_{j=1}^{k} \mu_j x_j \\
&\textstyle\sum_{j=1}^{k} c_j^1 x_j \leq C_0^1 \\
&\vdots \\
&\textstyle\sum_{j=1}^{k} c_j^L x_j \leq C_0^L \\
&\textstyle\sum_{j=1}^{k} x_j = 1 \\
&x_j \geq 0, \ \forall j
\end{aligned}
$$

A Linear Programming Formulation

$$
\begin{array}{lrl}
 & z_D(\underline{\theta}) & = \quad \min \; C_0^1 g_1 + \ldots + C_0^L g_L + g_{L+1} \\
 & & \quad c_1^1 g_1 + \ldots + c_1^L g_L + g_{L+1} \geq \mu_1 \\
\textbf{Dual} & & \qquad \vdots \\
 & & \quad c_k^1 g_1 + \ldots + c_k^L g_L + g_{L+1} \geq \mu_k \\
 & & \quad g_{L+1} \in \mathbf{R}, g_j \geq 0, \; j = 1, \ldots, L.
\end{array}
$$

Optimality Condition: $\phi_m \geq 0, m = 1, \ldots, k,$ where

$$\phi_m^B(\underline{\theta}) \equiv c_m^1 g_1^B + \ldots + c_m^L g_L^B + g_{L+1}^B - \mu_m \geq 0.$$

A basic matrix B satisfying this condition is optimal.

Model Description

- Sequential Sampling from k Statistical Populations
- Each period population i is sampled from:
  - X_i=random reward $\sim f_i(\cdot|\theta_i)$ (p.d.f. or p.m.f)
  - $\mu_i = \mathsf{E}(X_i|\theta_i)$
  - $\underline{\theta} = (\theta_1, \ldots, \theta_k)$
  - c_i=sampling cost
  - C_0=cost budget (on average basis)
- Objective: Identify a sampling policy for which
  - The long run expected average cost per period does not exceed C_0.
  - The long run expected average reward per sample is maximized.

A Linear Programming Formulation

x_j=Fraction of time periods allocated to sampling from population j

Primal

$$
\begin{aligned}
z(\underline{\theta}) = \max \quad & \textstyle\sum_{j=1}^{k} \mu_j x_j & \\
& \textstyle\sum_{j=1}^{k} c_j x_j & \leq C_0 \\
& \textstyle\sum_{j=1}^{k} x_j & = 1 \\
& x_j \geq 0, \ \forall j &
\end{aligned}
$$

Dual

$$
\begin{aligned}
z(\underline{\theta}) = \min \quad & C_0 w + v & \\
& c_j w + v & \geq \mu_j, \ j = 1, \ldots, k \\
& w \geq 0, & v \in \boldsymbol{R}
\end{aligned}
$$

Optimality Conditions

Two Cases for Optimal Solution

Case 1 $B = \begin{pmatrix} c_i & c_j \\ 1 & 1 \end{pmatrix}$, for some $i, j,\ c_j < C_0 < c_i$.

In this case $x_i = \frac{C_0 - c_j}{c_i - c_j},\ x_j = \frac{c_i - C_0}{c_i - c_j},\ x_m = 0, m \neq i, j.$

Dual Solution $w = \frac{\mu_i - \mu_j}{c_i - c_j},\ v = \frac{c_i \mu_j - c_j \mu_i}{c_i - c_j}.$

Optimality Condition: $\phi_m \geq 0, m = 1, \ldots, k,$ where

$$\phi_m = c_m w + v - \mu_m = \mu_i + (\mu_i - \mu_j)\frac{c_m - c_i}{c_i - c_j} - \mu_m$$

Optimality Conditions

Case 2 $B = \begin{pmatrix} c_i & 1 \\ 1 & 0 \end{pmatrix}$, for some i, $c_i \leq C_0$.

In this case $x_i = 1,\ x_m = 0, m \neq i$.

Dual Solution $w = 0,\ v = \mu_i$.

Optimality Condition: $\phi_m \geq 0, m = 1, \ldots, k,$ where

$$\phi_m = c_m w + v - \mu_m = \mu_i - \mu_m$$

The Incomplete Information Framework

Recall that the rewards $X_i \sim f_i(\cdot|\theta_i), i = 1, \ldots, k$.
We assume that

- $\theta_i, i = 1, \ldots, k$ are **unknown parameters**
- $\theta_i \in \Theta_i, \quad \underline{\theta} \in \Theta = \prod_{i=1}^{k} \Theta_i$
- Multi-Armed Bandit Framework

If no cost constraint were imposed, we would have a standard multi-armed bandit model with asymptotic average reward criterion.
Robbins(1952), Lai and Robbins (1985), Agrawal, Teneketzis and Anantharam (1989), Burnetas and Katehakis (1996), Kulkarni and Lugosi (2000), etc.

In the presence of side-constraints more work is required.

Adaptive Allocation Policies

When $\underline{\theta}$ is unknown, a sampling policy that uses this information is not admissible. Only **adaptive policies** are allowed.
Let A_t, Y_t be the population selected and the observed outcome at period t. Let $h_t = (a_1, x_1, \ldots, a_{t-1}, x_{t-1})$ be the history of actions and observations available at period t.
An **adaptive policy** is defined as a sequence $\pi = (\pi_1, \pi_2, \ldots)$ of history-dependent probability distributions on $\{1, \ldots, k\}$, so that

$$\pi_t(j, h_t) = \mathrm{P}(A_t = j | h_t)$$

Performance Measures

Define

- $T_n(j) = \sum_{t=1}^{n} 1(A_t = j)$=number of samples from j in first n periods
- $S_n^{\pi}(\underline{\theta}) = \mathrm{E}_{\underline{\theta}}^{\pi}(\sum_{t=1}^{n} Y_t) = \sum_{j=1}^{k} \mu_j(\theta_j)\mathrm{E}_{\underline{\theta}}^{\pi} T_n(j)$
  Expected $n-$period reward
- $C_n^{\pi}(\underline{\theta}) = \mathrm{E}_{\underline{\theta}}^{\pi}(\sum_{t=1}^{n} c_{A_t}) = \sum_{j=1}^{k} c_j \mathrm{E}_{\underline{\theta}}^{\pi} T_n(j)$
  Expected $n-$period cost

Definition

A policy π is called

- **feasible** ($\pi \in \Pi^F$), if $\limsup_{n\to\infty} \frac{C_n^{\pi}(\underline{\theta})}{n} \leq C_0, \ \forall \underline{\theta} \in \Theta$
- **consistent** ($\pi \in \Pi^C$), if $\pi \in \Pi^F$ and $\lim_{n\to\infty} \frac{S_n^{\pi}(\underline{\theta})}{n} = z(\theta), \ \forall \underline{\theta} \in \Theta$

Consistent Policies

Do consistent policies exist?
Yes, under very mild conditions. Construction is easy, one way is by using forced selections at sparse sequences of time periods (Robbins, 1952).

Assume that there exist consistent estimators $\hat{\mu}_{j,T_j}$ of $\mu_j(\theta_j)$. Let $\underline{\hat{\mu}}_n = (\hat{\mu}_{j,T_j(n)}, j = 1, \ldots, k)$.

Define the **certainty-equivalence LP** : $\hat{z}_n = z(\underline{\hat{\mu}}_n)$ and $\underline{\hat{x}}_n$ the corresponding primal solution : $\hat{z}_n = \underline{\hat{\mu}}_n^{'} \underline{\hat{x}}_n$.

$\underline{\hat{x}}_n$ suggests a sampling policy for period n: Randomize among $\{1, \ldots, k\}$ according to distribution $\underline{\hat{x}}_n$.

Consistent Policies

To ensure that $\lim_{n\to\infty} \underline{\hat{\mu}}_n = \underline{\mu}(\underline{\theta})$, consider k nonoverlapping sparse sequence of positive integers

$$\tau_j = \{\tau_{j,m},\ m = 1, 2, \ldots\},\ j = 1, \ldots, k,$$

such that $\lim_{m\to\infty} \frac{\tau_{j,m}}{m} = 0$.

Define a policy π^0 which in period n:

- If $n = \tau_{j,m}$ for some j, m then it selects $A_n = j$.
- Otherwise it follows the randomization suggested by the certainty equivalence solution $\underline{\hat{x}}_n$.

Theorem

$\pi^0 \in \Pi^C$.

A stricter form of the constraint

- A constraint of this type is not very realistic, since it allows the average cost to be above C_0 for arbitrarily long periods of time.
- We now impose a stricter form of the cost constraint, which requires the average sampling cost not to exceed C_0 at any intermediate step and not only in the limit.
- Equivalently, assume that the experimenter has a sampling budget S.
  - At the beginning of each period a a fixed amount C_0 is added to the budget.
  - Also in each period, the experimenter can select any population to sample from, as long as the sampling cost does not exceed the budget.
  - If the budget is S_n in period n, then we can sample from any population i such that $c_i < S_n + C_0$. In this case the budget left for the next period is $S_{n+1} = S_n + C_0 - c_i$.

Consistency under the Stricter Constraint

- Consistent policies can be constructed under the new stricter constraint framework.
- Generalize the idea of forced sampling along a sparse sequence of time points
- Self-financing Sampling Block: A sequence of time periods in which
  1. Sampling is exclusively performed from a specific group of two or more populations.
  2. The sampling budget is zero at the beginning and at the end of the block.
  3. The budget is never violated.
- Replace sampling periods by sampling blocks.

Comparison of Consistent Policies

If there are many consistent policies, how can we select the "best" among them? Convergence rate, Loss Function.

A Loss Function

$L_n^{\pi}(\underline{\theta})$ = Expected loss from optimal policy performance in n periods, when adaptive policy π is followed and the true parameter value is $\underline{\theta}$.

When there is no constraint the loss function takes a simple form

$$\begin{aligned} L_n^{\pi}(\underline{\theta}) &= n\mu^*(\underline{\theta}) - S_n^{\pi}(\underline{\theta}) = n\mu^*(\underline{\theta}) - \sum_{j=1}^{k} \mu_j(\theta_j) \mathrm{E}_{\underline{\theta}}^{\pi} T_n(j) \\ &= \sum_{j=1}^{k} (\mu^*(\underline{\theta}) - \mu_j(\theta_j)) \mathrm{E}_{\underline{\theta}}^{\pi} T_n(j) \end{aligned}$$

Obviously $L_n^{\pi}(\underline{\theta}) \geq 0, \ \forall \pi, \underline{\theta}$.

A Loss Function

For the constrained problem

$$L_n^{\pi}(\underline{\theta}) = nz(\underline{\theta}) - S_n^{\pi}(\underline{\theta})$$

and $L_n^{\pi}(\underline{\theta}) \geq 0$ for $\pi \in \Pi^F$.
The following result relates the loss function with the sampling frequencies from the nonoptimal populations and it is useful for characterizing efficient policies. It follows easily from linear programming duality.
Recall that

- $\phi_j(\underline{\theta}) = C_j w(\underline{\theta}) + v(\underline{\theta}) - \mu_j(\theta_j)$: Optimality test quantity for population j. For optimal populations $\phi_j = 0$, for nonoptimal $\phi_j > 0$.
- $S_n^{\pi}(\underline{\theta})$=expected n-period reward for policy π.
- $C_n^{\pi}(\underline{\theta})$=expected n-period cost for policy π.

Loss Function Decomposition

Proposition

For $\pi \in \Pi^F$,

$$L_n^\pi(\underline{\theta}) = \sum_{j=1}^{k} \phi_j(\underline{\theta}) \mathrm{E}_{\underline{\theta}}^\pi T_n(j) + w(nC_0 - C_n^\pi(\underline{\theta})).$$

The proposition decomposes the optimality and the feasibility effects.

- $\sum_{j=1}^{k} \phi_j(\underline{\theta}) \mathrm{E}_{\underline{\theta}}^\pi T_n(j) =$ Loss due to sampling from nonoptimal populations (optimality effect).
- $w(nC_0 - C_n^\pi(\underline{\theta})) =$ Loss due to sampling from the optimal populations with incorrect frequencies (feasibility effect).

Policy Refinement

For a consistent policy, $\forall \underline{\theta} \in \Theta$,

$$\lim_{n\to\infty} \frac{S_n^\pi(\underline{\theta})}{n} = z(\theta)$$

thus

$$\lim_{n\to\infty} \frac{L_n^\pi(\underline{\theta})}{n} = 0, \quad \text{i.e.,} \quad L_n^\pi(\underline{\theta}) = o(n).$$

The following is a refinement of the consistency property.

Definition

A policy π is uniformly fast ($\pi \in \Pi^F$) if

$$L_n^\pi(\underline{\theta}) = o(n^\alpha), \forall \alpha > 0, \underline{\theta} \in \Theta.$$

Do uniformly fast policies exist? If they do, is there an "optimal" policy in this class?

An Asymptotic Lower Bound on the Loss Function

Theorem

There exist $K_j(\theta_j), j = 1, \ldots, k$ such that $\forall \pi \in \Pi^F, \underline{\theta} \in \Theta$

$$\liminf_{n\to\infty} \frac{\mathrm{E}_{\underline{\theta}}^{\pi} T_n(j)}{\log n} \geq \frac{1}{K_j(\theta_j)}$$

Therefore,

$$\liminf_{n\to\infty} \frac{L_n^{\pi}(\underline{\theta})}{\log n} \geq M(\underline{\theta}) \equiv \sum_{j=1}^{k} \frac{\phi_j(\underline{\theta})}{K_j(\theta_j)}.$$

If a policy π induces a loss that grows with rate lower than any polynomial function for all possible parameter configurations, then for the same policy the loss must grow with at least logarithmic rate.

An Asymptotically Optimal Policy

Theorem

There exists an adaptive allocation policy π^* such that $\underline{\theta} \in \Theta$

$$\limsup_{n\to\infty} \frac{L_n^{\pi}(\underline{\theta})}{\log n} \leq M(\underline{\theta}).$$

Policy π^* is **asymptotically optimal**.

Optimal in the class of policies that have a low loss function uniformly in the parameter $\underline{\theta}$.

The Constant $M(\underline{\theta})$

$$M(\underline{\theta}) = \sum_{j=1}^{k} \frac{\phi_j(\underline{\theta})}{K_j(\theta_j)}$$

where

$$K_j(\theta_j) = \inf\{I(\theta_j, \theta_j') : \theta_j' \in \Theta_j, \phi_j(\theta_j') < 0\},$$

and $I(\theta_i, \theta_i')$ is the Kullback-Leibler information (cross-entropy) between $f_i(\cdot|\theta_i)$ and $f_i(\cdot|\theta_i')$

$$I(\theta_i, \theta_i') = \mathrm{E}_{\theta_i}\left(\log \frac{f_i(X_i|\theta_i)}{f_i(X_i|\theta_i')}\right) = \sum_l \theta_{il} \log \frac{\theta_{il}}{\theta_{il}'}$$

Construction of an Asymptotically Optimal Policy

To define an asymptotically optimal policy we use SB of two types : Initial Sampling and Linear Programming

- An IS-Block (ISB) is to have estimates for all the populations in a self-financing way.
- A LP-Block (LPB) is to imitate the randomization suggested by certainty equivalence in a self-financing way.

Note that here we do not count time periods but number of blocks.

Definition of π^*

Policy π^* employs sampling blocks of the two types.
The first block ($r = 1$) is an ISB. Then, the r-th sampling block, that starts in period n is obtained as follows

- at the end of (r-1)-th block we have estimates $\underline{\hat{\theta}}^r$ with $\mu_1(\hat{\underline{\theta}}_1^r), ..., \mu_k(\hat{\underline{\theta}}_k^r)$ which gives the solution $z(\underline{\hat{\theta}}^r)$
- we inflate the estimates $\mu_\alpha(\hat{\underline{\theta}}_\alpha^r)$ using :
  $u_\alpha(\hat{\underline{\theta}}_\alpha^r, \gamma) = \sup_{\hat{\underline{\theta}}_\alpha^{r'}} \{\mu_\alpha(\hat{\underline{\theta}}_\alpha^{r'}) : I(\hat{\underline{\theta}}_\alpha^r, \hat{\underline{\theta}}_\alpha^{r'}) < \gamma\}$
- if $u_\alpha(\hat{\underline{\theta}}_\alpha^r, \gamma) > \mu_\alpha(\hat{\underline{\theta}}_\alpha^r) + \phi(\underline{\hat{\theta}}^r)$ for some population α, let $z_\alpha(\underline{\hat{\theta}}^r)$ be the LP solution where $\mu_\alpha(\hat{\underline{\theta}}_\alpha^r)$ has been replaced by $u_\alpha(\hat{\underline{\theta}}_\alpha^r, \gamma)$ for population α only
- let $z^*(\underline{\hat{\theta}}^r) = max_\alpha z_\alpha(\underline{\hat{\theta}}^r)$
- employ the LPB which corresponds to the LP solution $z_\alpha(\underline{\hat{\theta}}^r)$ that gives the maximum value $z^*(\underline{\hat{\theta}}^r)$.

Note that if all populations have $u_\alpha(\hat{\underline{\theta}}_\alpha^r, \gamma) < \mu_\alpha(\hat{\underline{\theta}}_\alpha^r) + \phi(\underline{\hat{\theta}}^r)$ then we solve the certainty equivalence LP and employ the corresponding LPB.

Normal Distributions with Unknown Means and Known Variances

Assume the observations X_α^j from bandit α are normally distributed with unknown means $EX_\alpha^j = \theta_\alpha$ and known variances σ_α^2, i.e., $\underline{\theta}_\alpha = \theta_\alpha$, $\mu_\alpha(\underline{\theta}_\alpha) = \theta_\alpha$. Given history h_l, define

$$\mu_\alpha(\hat{\theta}_\alpha^l) = \frac{\sum_{j=1}^{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))} X_\alpha^j}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))}.$$

Also, we have:

$$I(\theta_\alpha, \theta_\alpha^{'}) = \frac{(\theta_\alpha^{'} - \theta_\alpha)^2}{2\sigma_\alpha^2}$$

$$K_\alpha(\underline{\theta}) = \frac{(\phi_\alpha^B(\underline{\theta}))^2}{2\sigma_\alpha^2}.$$

Normal Distributions with Unknown Means and Known Variances

It is easy to see that the indices simplify to:

$$\theta'_\alpha = \hat{\theta}^I_\alpha + \sigma_\alpha \left(\frac{2 \log S_{\pi^0}(I-1)}{T^\alpha_{\pi^0}(S_{\pi^0}(I-1))} \right)^{1/2}.$$

is the θ'_α which satisfies the supremum in the index $u_\alpha(\underline{\hat{\theta}}^I)$.

Normal Distributions with Unknown Means and Unknown Variances

Assume the observations X_α^j from bandit α are normally distributed with unknown means $EX_\alpha^j = \mu_\alpha$ and unknown variances $VarX_\alpha^j = \sigma_\alpha^2$, i.e., $\underline{\theta}_\alpha = (\mu_\alpha, \sigma_\alpha^2)$. Given history h_l, define

$$\mu_\alpha(\hat{\underline{\theta}}_\alpha^l) = \frac{\sum_{j=1}^{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))} X_\alpha^j}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))}$$

and

$$\sigma_\alpha^2(\hat{\underline{\theta}}_\alpha^l) = \frac{\sum_{j=1}^{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))} (X_\alpha^j - \mu_\alpha(\hat{\underline{\theta}}_\alpha^l))^2}{T_{\pi^0}^\alpha(S_{\pi^0}(l-1))}.$$

Normal Distributions with Unknown Means and Unknown Variances

Also, we have:

$$I(\underline{\theta}_\alpha, \underline{\theta}'_\alpha) = \frac{(\mu_\alpha(\underline{\theta}'_\alpha) - \mu_\alpha)^2 + \sigma_\alpha^2 - \sigma_\alpha^2(\underline{\theta}'_\alpha)}{2\sigma_\alpha^2(\underline{\theta}'_\alpha)} + \log \frac{\sigma_\alpha(\underline{\theta}'_\alpha)}{\sigma_\alpha}$$

$$K_\alpha(\underline{\theta}) = \frac{1}{2} \log \left(1 + \frac{(\phi_\alpha^B(\underline{\theta}))^2}{\sigma_\alpha^2} \right).$$

It is easy to see that the UCB indices simplify to:

$$\mu_\alpha(\underline{\theta}'_\alpha) = \mu_\alpha(\hat{\underline{\theta}}^I_\alpha) + \sigma_\alpha(\hat{\underline{\theta}}^I_\alpha) \left(S_{\pi^0}(I-1)^{\frac{2}{T^\alpha_{\pi^0}(S_{\pi^0}(I-1))-2}} - 1 \right)^{1/2}$$

and

$$\sigma_\alpha^2(\underline{\theta}'_\alpha) = (\mu_\alpha(\underline{\theta}'_\alpha) - \mu_\alpha(\hat{\underline{\theta}}^I_\alpha))^2 + \sigma_\alpha^2(\hat{\underline{\theta}}^I_\alpha),$$

are the mean and variance of the $\underline{\theta}'_\alpha$ which satisfy the supremum in the index $u_\alpha(\hat{\underline{\theta}}^I)$.

Future Work

- Bayesian learning for constrained MABs
- Inventory models with bayesian learning using MDPs.
- Adaptive policies for a two-product inventory model with dynamic customers.
- Healthcare models using MDPs: A Model of Managing Chronic Care with Patient Activation Measure.
- Call-Back Option via Dynamic Prioritization in a Call Center.

Literature

Bayesian Approach of MAB Problem with Constraints

- S. Dayanik, W. Powell, and K. Yamazaki (2008). Index policies for discounted bandit problems with availability constraints. Advances in Applied Probability, 40(2) pp 377-400.
- E. Denardo, E. Feinberg, and U. Rothblum (2013). The multi-armed bandit with constraints. Annals of Operations Research, 208(1) pp 37-62.
- Cowan W. and M.N. Katehakis (2015). Multi-armed Bandits under General Depreciation and Commitment, Probability in the Engineering and Informational Sciences, 29 (1) pp 51-76.

Literature

Frequentist Approach of MAB Problem with Constraints

- L. Tran-Thanh, A. Chapman, Munoz De Cote Flores Luna, J. Enrique, A. Rogers, and N. R Jennings (2010). Epsilon-first policies for budget-limited multi-armed bandits. In AAAI-10, pp 1211-1216.
- A. N. Burnetas and O. A. Kanavetas (2012). Adaptive policies for sequential sampling under incomplete information and a cost constraint. N. J. Daras (ed.) Application of Mathematics and Informatics in Military Science, Springer, pp 97-112.
- A. Badanidiyuru, R. Kleinberg, and A. Slivkins (2013). Bandits with knapsacks. In Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium, pp 207-216.
- Burnetas, A. N., Kanavetas, O. A. and M. N. Katehakis (2017). Asymptotically Optimal Multi-Armed Bandit Policies under a Cost Constraint. Probability in the Engineering and Informational Sciences, 31 (3), pp. 284-310.
- Burnetas, A. N., Kanavetas, O. A. and M. N. Katehakis (2018). Optimal Data Driven Resource Allocation under Multi-Armed Bandit Observations. *Under revision in Management Science*
- Burnetas, A. N., Kanavetas, O. A. and M. N. Katehakis (2019). Asymptotically Optimal Control for Markov Decision Processes (MDP) under Side Constraints. *Working paper*