Algorithms for Pure Exploration

Source: Bandit Algorithms, L. & S., Chap 33.

Setting: k-arms, unstructured environment $\nu \in \mathcal{E}$.

Goal: Pure exploration, i.e., $\pi = (\pi_t, t = 1, 2, \ldots, n+1)$ a policy (mapping from history to an arm index).

i.e., $A_1, X_1, A_2, X_2, \ldots, A_n, X_n, A_{n+1}$

- A_1: action = arm play index at time = 1
- X_1: reward at $t = 1$
- A_n: action at time = n
- X_n: reward at time = n
- A_{n+1}: the action at time $(n+1)$ — regret depends ONLY on this action

Simple Regret: $R_n^{\text{simple}}(\pi, \nu) = \mathbb{E}_{\nu\pi}\left[\Delta_{A_{n+1}}(\nu)\right]$

- n: time steps provided to "experiment"
- π: policy
- ν: environment

$\Delta_i(\nu)$ = gap of i^{th} arm wrt best arm in environment ν, $\nu \in \mathcal{E}$.

The Simplest Algorithm: Uniform Explore.

Input: n, k

Iteration: for each $t = 1, 2, \dots, n$:

$$A_t = 1 + t \bmod k$$

Output (at time $n+1$): $A_{n+1} = \underset{1 \le i \le k}{\arg\max}\ \hat{\mu}_i(n)$

where $\hat{\mu}_i(n)$ = sample average of i^{th} arm rewards.

Thm (33.1 in textbook): $\mathcal{E} = \mathcal{E}_{SG}^k(1)$, $\nu \in \mathcal{E}_{SG}^k(1)$, an environment (1-subG). For all $n \ge k$,

$$R_n^{\text{simple}}(\pi, \nu) \le \min_{\Delta \ge 0} \left\{ \Delta + \sum_{\{i : \Delta_i(\nu) > \Delta\}} \Delta_i(\nu)\, e^{\frac{-\lfloor \frac{n}{k} \rfloor \Delta_i(\nu)^2}{4}} \right\}$$

Proof: follows immediately from 1-subG. For each arm, we have atleast $\lfloor \frac{n}{k} \rfloor$ samples. Assume (wLOG) that arm 1 is optimal (i.e. $\Delta_1 = 0$).

For any i (sub-optimal arm) with $\Delta_i > 0$, we have:

$$P\left(\hat{\mu}_i(n) \geq \hat{\mu}_1(n)\right) = P\left(\underbrace{\hat{\mu}_i(n) - \hat{\mu}_1(n)}_{\sqrt{2}\text{-subG with mean } -\Delta_i} \geq 0\right)$$

$$\leq e^{-\frac{\lfloor \frac{n}{k} \rfloor \Delta_i^2}{4}}$$

$$\therefore R_n^{\text{Simple}}(\pi, \nu) = \mathbb{E}_{\nu\pi}\left[\Delta_{A_{n+1}}(\nu)\right]$$

$$= \sum_{i=2}^{k} \Delta_i P(A_{n+1} = i)$$

(sub optimal arms)

Now, choose any $\Delta > 0$, and let $S = \{i \leq k : \Delta_i > \Delta\}$

$$= \sum_{i \notin S} \Delta_i P(A_{n+1} = i) + \sum_{i \in S} \Delta_i P(A_{n+1} = i)$$

$$\leq \Delta + \sum_{i \in S} \Delta_i e^{-\frac{\lfloor \frac{n}{k} \rfloor \Delta_i^2}{4}}$$

$\square$

Note 1: For fixed k, fixed environment $\nu \in \mathcal{E}_{SG}^k(1)$, the simple regret decays exponentially in time horizon n. In other words, the prob of not playing the optimal arm at time $(n+1)$ is exponentially small..

Note 2: (Worst-case bound over all environments $\nu \in \mathcal{E}^k_{SG}(1)$)

Use B-H inequality to show that $\exists\, \nu \in \mathcal{E}^k_{SG}(1)$ s.t.

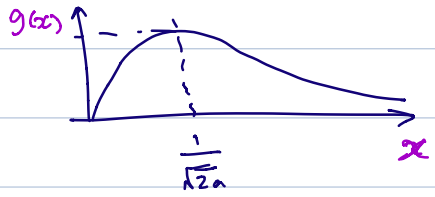
$$R_n^{simple}(\pi, \nu) \geq C\sqrt{k/n}$$

With above Uniform Explore policy, recall that

$$R_n^{simple}(\pi, \nu) \leq \min_{\Delta \geq 0} \left\{ \Delta + \sum_{\{i : \Delta_i(\nu) > \Delta\}} \Delta_i(\nu)\, e^{-\frac{n}{\lfloor k \rfloor} \frac{\Delta_i^2(\nu)}{4}} \right\}$$

Choose $\Delta = 2\sqrt{\dfrac{\ln(k)}{\lfloor n/k \rfloor}}$, and assume $n \geq k$.

Aside: $z^* = \max_{x \geq b} \underbrace{x e^{-ax^2}}_{g(x)}$



$\therefore\ b > \frac{1}{\sqrt{2a}} \Rightarrow z^* = g(b)$

$$R_n^{simple}(\pi, \nu) \leq 2\sqrt{\frac{\ln(k)}{\lfloor n/k \rfloor}} \left(1 + (k-1)e^{-\ln k}\right)$$

$$\leq 4\sqrt{\frac{\ln(k)}{\lfloor n/k \rfloor}} \leq 4\sqrt{\frac{k \ln k}{n}}$$

$\Rightarrow$ Extra $\sqrt{\ln k}$ factor compared to lower bound.

An alternate policy that (in the worst case) gets rid of the $\sqrt{\ln(k)}$

Let π be any policy that optimizes for cumulative regret (e.g. UCB). Let $T_j(t)$ be the number of times that arm $j \in \{1, 2, \ldots, k\}$ is played under this policy π.

Now define the policy $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2 \cdots \hat{\pi}_n \hat{\pi}_{n+1})$, with

$$\hat{\pi}_t = \pi_t \quad , \quad t = 1, 2, \ldots, n$$

and $\hat{\pi}_{n+1} \equiv \hat{\pi}_{n+1}(j \mid a_1, x_1, a_2, x_2, \ldots, a_n, x_n)$

$$= \frac{T_j(n)}{n} \quad , \quad j = 1, 2, \ldots, k$$

(aka "choose best arm proportional to the relative frequency among played arms").

Prop (§3.2 in text): The simple regret of this policy $\hat{\pi} = (\hat{\pi}_t, t = 1, 2, \ldots, n, n+1)$ is

Given by: $R_n^{\text{simple}}(\hat{\pi}, \nu) = \frac{R_n(\pi, \nu)}{n}$

cumulative regret of policy π.

Proof: Regret decomposition implies

$$R_n(\pi, \nu) = \sum_{i=1}^{k} \Delta_i \, E[T_i(n)]$$

$$= n \, E\left[\sum_{i=1}^{k} \Delta_i \frac{T_i(n)}{n} \right]$$

$$= n \, E[\Delta_{A_{n+1}}] = n \, R_n^{\text{simple}}(\hat{\pi}, \nu). \quad \blacksquare$$

Recall that $\exists \pi$ s.t. $R_n(\pi, \nu) \leq C\sqrt{nk}$ for all $\nu \in \mathcal{E}_{SG}^{k}(1)$, from earlier discussion on worst-case regret bounds.

$$\Rightarrow R_n^{\text{simple}}(\hat{\pi}, \nu) \leq C\sqrt{\frac{k}{n}}$$

Note 3: The above does NOT mean that $\hat{\pi}$ is a good policy in general. The reason

is that in a fixed environment (not worst case), the above policy $\hat{\pi}$ provides only a polynomial decay with n (time horizon), whereas the Uniform Explore policy provides exponential decay.

Please read Sec. 33.1 for more discussion.

Fixed Confidence Level $\delta \in (0,1)$

Goal: Identify the best arm, with a prespecified error prob. The number of samples is a free variable, and needs to be minimized.

Formally, let $\mathcal{F}_t = \sigma\{A_1, X_1, \ldots, A_t, X_t\}$, i.e., the σ-algebra corresponding to the information until time t.

Let τ be a stopping time wrt the filtration $\{\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \ldots\}$. Informally, τ is a "causal" r.v., i.e., for each $t \geq 1$, there exists a function $g_t(\cdot)$ such that:

$$\{\tau \le t\} = \{ g_t(A_1, X_1, \cdots, A_t, X_t) \le t \}$$

A Player determines a triplet: (π, τ, ψ) → decision

arm play policy ← π — τ ↓ Stopping time (when to stop playing)

$\pi = (\pi_1, \pi_2 \cdots)$

$\pi_t : (A_1, X_1, \cdots, A_{t-1}, X_{t-1}) \longmapsto \{1, 2, \cdots, k\}$

① $\pi = (\pi_1, \pi_2, \pi_3, \pi_4 \cdots \pi_t \cdots)$,

history of past observations

where $\pi_t : \mathcal{H}_{t-1} \longmapsto \{1, 2, \cdots, k\}$

② τ a stopping time, determines when to stop playing and declare that a specific arm to be the best arm.

③ $\psi \in \{1, 2, \cdots, k\}$ is the selection rule, and chooses which arm to select as the best arm.

Sound Triplet (π, τ, ψ) A triplet is said to be sound at some pre-selected confidence level $\delta \in (0, 1)$ if

$$P_{\nu\pi}(\{\tau < \infty\} \cap \{\Delta_\psi(\nu) > 0\}) \le \delta$$

In other words, a sound policy that stops in a finite time does not select a sub-optimal arm with probability exceeding δ.

Note 4: With multiple optimal arms, a sound policy will not be able to stop in finite time. (impossible to distinguish between noise and arms that are arbitrarily close); see comment in Sec. 33.2.

Lower Bound

$\mathcal{E}^k$ an arbitrary k-armed stochastic bandit; $\nu \in \mathcal{E}^k$.

$$i^*(\nu) = \underset{1 \le j \le k}{\operatorname{argmax}} \; \mu_j(\nu)$$

↳ mean reward of i^{th} arm in environment $\nu \in \mathcal{E}^k$.

$$\mathcal{E}^k_{alt}(\nu) = \left\{ \nu' \in \mathcal{E} : i(\nu') \cap i(\nu) = \emptyset \right\}$$

↳ $\subseteq \mathcal{E}$ s.t. optimal arm in ν' is different from an optimal in ν.

Recall Divergence Decomposition Thm with Stopping Times (HW problem 15.7 in text):

Thm: τ a stopping time with respect to filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 1\}$ where

$$\mathcal{F}_t = \sigma\{A_1, X_1, \cdots, A_t, X_t\}.$$

If Z is a r.v. that is $\mathcal{F}_\tau$-measurable, then

$$D\left(P_{\nu Z}, P_{\nu' Z}\right) \leq \sum_{j=1}^{k} E_\nu\left[T_j(\tau)\right] \cdot D\left(P_j, P_j'\right)$$

where $P_{\nu Z}$ is the prob law of Z under the environment ν, and analogous for $P_{\nu' Z}$.

For comparison, see Set 11 / Lemma 15.1 in text. There, $Z \longrightarrow X$, and $\tau \longrightarrow t$. The result showed a seperation between policy π and arm divergence. A similar result holds (above) even with stopping times.

Note: $\mathcal{F}_\tau$ is the σ-algebra at a stopping time τ,

defined as follows: $\mathcal{F} = \{\mathcal{F}_t, t=1,2,\dots\}$.

$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \le t\} \in \mathcal{F}_t \quad \forall t=1,2,\dots\}$$

This is indeed a valid σ-algebra, and intuitively, this describes the "information" at the random time τ.

Theorem (33.5 in text): Suppose (π, τ, ψ) is sound for $\mathcal{E}^k$ at confidence level $\delta \in (0,1)$. Further, let $\nu \in \mathcal{E}$. Then,

$$E_{\nu\pi}[\tau] \ge c^*(\nu) \ln\left(\frac{1}{4\delta}\right), \text{ where}$$

$$c^*(\nu)^{-1} = \sup_{\alpha \in \mathcal{P}^{k-1}} \left(\inf_{\nu' \in \mathcal{E}^k_{alt}(\nu)} \sum_{i=1}^{k} \alpha_i(\nu) D(\nu_i; \nu_i') \right)$$

Proof: WLOG $E_{\nu\pi}[\tau] < \infty$ (else above immediately true). Thus τ is a proper r.v. (i.e. no mass at ∞).

Pick any $\nu' \in \mathcal{E}^*_{alt}$, and $B = \{\tau < \infty\} \cap \{\psi \notin i^*(\nu)\}$

i.e., B is a bad event in the ν' environment. Further,

$$B^c = \left(\{\tau < \infty\} \cap \{\psi \notin i^*(\nu')\}\right)^c$$

$$= \{\tau = \infty\} \cup \{\psi \in i^*(\nu')\}$$

$$= \{\psi \in i^*(\nu')\}$$

$$\subseteq \{\psi \notin i^*(\nu)\} \qquad \left(\because\ i(\nu) \cap i(\nu') = \emptyset \text{ from } \nu' \in \Sigma^k_{alt}\right)$$

$$= \{\psi \notin i^*(\nu)\} \cap \{\tau < \infty\} \qquad \left(\because\ \tau \text{ is a proper r.v.}\right)$$

$\hookrightarrow$ (1)

From soundness, we have:

$$P_{\nu\pi}\left(\{\tau < \infty\} \cap \{\psi \notin i^*(\nu)\}\right) \leq \delta$$

$$P_{\nu'\pi}\left(\{\tau < \infty\} \cap \{\psi \notin i^*(\nu)\right) \leq \delta$$

$\hookrightarrow$ (2)

Combining (1) and (2), we have:

$$2\delta \underset{(1)}{\geq} P_{\nu\pi}\left(\{\tau < \infty\} \cap \{\psi \notin i^*(\nu)\}\right) + P_{\nu'\pi}\left(\{\tau < \infty\} \cap \{\psi \notin i^*(\nu)\}\right)$$

$$\geq \underset{(2)}{} P_{\nu\pi}(B^c) + P_{\nu'\pi}(B)$$

$$\geq \frac{1}{2} e^{-D(P_{\nu\pi}, P_{\nu'\pi})}$$

B-H inequality (see Set 10; Chap 14 in text)

$$\geq \frac{1}{2} e^{-\sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)] D(\nu_j, \nu_j')}$$

Divergence Decomp with stopping time

→ (3)

∴ From (3), we have: For any $\nu' \in \mathcal{E}_{alt}^k$,

$$\sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)] D(\nu_j, \nu_j') \geq \ln\left(\frac{1}{4\delta}\right)$$

→ (4)

(Aside: this implies $E_{\nu\pi}[\tau] > 0$, thus $0 < \frac{1}{E_{\nu\pi}[\tau]} < \infty$).

Finally,

$$\frac{E_{\nu\pi}[\tau]}{C^*(\nu)} \overset{\text{defn. of } C^*(\nu)}{=} E_{\nu\pi}[\tau] \sup_{\alpha \in \rho^{d-1}} \inf_{\nu' \in \mathcal{E}_{alt}^k} \sum_{j=1}^{k} \alpha_j D(\nu_j, \nu_j')$$

$$\geq E_{\nu\pi}[\tau] \inf_{\nu' \in \mathcal{E}_{alt}^k} \sum_{j=1}^{k} \underbrace{\frac{E_{\nu\pi}[T_j(\tau)]}{E_{\nu\pi}[\tau]}}_{\alpha_i} D(\nu_j, \nu_j') \quad \text{s.t. } \sum_i \alpha_i = 1,\ \alpha_i \geq 0$$

$$= \inf_{\nu' \in \mathcal{E}^k_{alt}} \sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)] \, D(\nu_j, \nu_j')$$

$$\geq \ln\left(\frac{1}{4\delta}\right). \quad (4)$$

∎

Some intuition on $c^*(\nu)$

Fix an environment $\nu \in \mathcal{E}^k$, and let the gap be Δ between the best and second best arm. Recall that

$$c^*(\nu)^{-1} = \sup_{\alpha \in \mathcal{P}^{k-1}} \left(\inf_{\nu' \in \mathcal{E}^k_{alt}(\nu)} \sum_{j=1}^{k} \alpha_j D(\nu_j, \nu_j') \right)$$

Potential problem: For a fixed α, we might find a sequence of $\nu' \in \mathcal{E}^k_{alt}(\nu)$ s.t. $D(\nu_j, \nu_j') \to 0$, and thus making $c^*(\nu)^{-1} = 0$ for any $\nu \in \mathcal{E}^k$.

The above is not true. Consider as an example $\mathcal{E}^k = \mathcal{E}^2_N(1)$ (i.e., 2-armed 1-Gaussian bandit). Lets say $\nu = (\nu_1, \nu_2) = (N(\Delta, 1), N(0, 1))$, i.e., $\mu_1 = \Delta$, $\mu_2 = 0$.

$$D(\nu_j, \nu_j') = \frac{1}{2}(\mu_j - \mu_j')^2 \qquad \text{(for Gaussian r.v.s.)}$$

$\nu' \in \mathcal{E}^2_{alt} = \{(\mu_1', \mu_2') : \mu_2' > \mu_1'\}$

Suppose we choose $\mu' = (\mu_1', \mu_2')$ s.t. $\mu_2' = \mu_2 = 0$ (i.e., forcing the divergence for arm 2 to go to zero), the constraint that $\mu_1' < \mu_2' = 0$, and that $\mu_1 = \Delta > 0 \Rightarrow D(\mu_1, \mu_1') \geq \frac{\Delta^2}{2} > 0$.

More generally, we cannot construct an alternative environment ν' s.t. $D(\mu_j, \mu_j') \approx 0$ for all arms. Specifically for the case where

$$\nu = (N(\mu_1, 1) \quad N(\mu_2, 1)), \qquad \nu \in \mathcal{E}^2_N(1)$$

with $\mu_1 > \mu_2$, we have:

$$\mathcal{E}^2_{alt} = \{(N(\mu_1', 1) \quad N(\mu_2', 1)) : \mu_2' > \mu_1'\}$$

Then, $$C^*(\nu)^{-1} = \sup_{\alpha \in \mathcal{P}^{k-1}} \left(\inf_{\nu' \in \mathcal{E}^k_{alt}(\nu)} \sum_{j=1}^{k} \alpha_j D(\nu_j, \nu_j') \right)$$

$$= \sup_{\alpha \in [0,1]} \inf_{\nu' \in \mathcal{E}^2_{alt}} \alpha \frac{(\mu_1 - \mu_1')^2}{2} + (1-\alpha)\frac{(\mu_2 - \mu_2')^2}{2}$$

$$= \sup_{\alpha \in [0,1]} \frac{1}{2} \alpha (1-\alpha) (\mu_1 - \mu_2)^2 \qquad \text{(Check!)}$$

$$= \frac{1}{8} (\mu_1 - \mu_2)^2 \quad , \text{ with } \alpha^* = \frac{1}{2}$$

(See textbook, sec. 33.2.1 for a more extensive discussion of $c^*(\nu)$)

The Track and Stop Algorithm

Goal: Algorithm that asymptotically meets the lower bound: (π, τ, ψ) s.t. lower bound met as $\delta \to 0$.

Setting: $\mathcal{E} = \mathcal{E}_N^k(1)$, i.e., rewards are Gaussian with $\sigma^2 = 1$ for each arm.

Intuition: We need an algorithm (π, τ, ψ) s.t.

$$E_{\nu \pi}[\tau] \underset{\text{as } \delta \to 0}{\gtrsim} c^*(\nu) \cdot \ln\left(\frac{4}{\delta}\right).$$

Now recall the lower bound proof. Towards the end,

we had the following chain:

$$\frac{E_{\nu\pi}[\tau]}{C^*(\nu)} \geq \inf_{\nu' \in \mathcal{E}^k_{alt}} \sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)]\, D(\nu_j, \nu_j')$$

$$\geq \ln\left(\frac{1}{\delta}\right).$$

If we hope that this is tight (at least in an asymptotic in δ sense), then, we should expect that a "good" (π, τ, ψ) triplet should satisfy: $\ln\left(\frac{1}{4\delta}\right) + \ln\left(\frac{1}{\delta}\right)$

$$\inf_{\nu' \in \mathcal{E}^k_{alt}} \sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)]\, D(\nu_j, \nu_j') \approx \ln\left(\frac{1}{4\delta}\right) \quad \hookrightarrow (1)$$

In the regime where $\delta \to 0$, we expect $\tau \to \infty$, and $T_j(\tau) \to \infty$. To get asymptotic optimality, we thus have some slack: Instead of hoping to satisfy (1), we can try to satisfy:

$$\inf_{\nu' \in \mathcal{E}^k_{alt}} \sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)]\, D(\nu_j, \nu_j') \approx h(t) + \ln\left(\frac{1}{\delta}\right) \quad \hookrightarrow (2)$$

$h(t)$: some slowly growing function of t s.t. this term is not dominant, but nevertheless gives flexibility to ensure sufficient concentrations.

In a Gaussian setting $\left(\mathcal{E}_N^k(1)\right)$, $D(\nu_i, \nu_i') = \frac{(\mu_i - \mu_i')^2}{2}$

(a) Thus, we are looking for a criteria that uses the above quadratic.

(b) Further, since we do not have access to $\{\mu_j, j=1,2,\dots k\}$, and we are on a single sample path (i.e. no $E_{\nu\pi}[\cdot]$), we can hope to replace the appropriate parameters by a sample average and/or single sample.

(c) Recall in the lower bound that $\alpha_j(\nu) = \frac{E_{\nu\pi}[T_j(\tau)]}{E_{\nu\pi}[\tau]}$,

$j=1,2,\dots,k$. Thus, the intuition is that this is the fraction of time to spend sampling arm j.

Motivated by above, three ideas to develop an algorithm, characterized by (π, τ, ψ):

(1) Choosing τ: In (2) above, replace $E_{\pi\nu}[\cdot]$ by the sample (i.e., get rid of $E[\cdot]$), and replace μ_j by $\hat{\mu}_j(t)$ = empirical estimate of arm j's mean at time t. Finally, estimate best current

arm using the empirical averages (i.e., no UCB or any variance correction). Then define a surrogate for LHS in (2) as Z_t, where:

$$Z_t = \inf_{\nu' \in \mathcal{E}^k_{alt}(\hat{\nu}(t))} \sum_{j=1}^{k} T_j(t) \frac{(\hat{\mu}_j(t) - \mu'_j)^2}{2}$$

- $\nu' \in \mathcal{E}^k_{alt}(\hat{\nu}(t))$: i.e., avoid the currently estimated best arm.
- $T_j(t)$: use sample instead of $E_{\pi\nu}[\cdot]$
- $\hat{\mu}_j(t) - \mu'_j$: use empirical estimate to replace unknown μ_j

Here, there is a unique best arm, defined by $i^*(t) = \arg\max_j \hat{\mu}_j(t)$. If several $\hat{\mu}_j(t)$ are equal, pick any one of them to be the (sole) declared best arm.

Original Expression in (2)

$$\inf_{\nu' \in \mathcal{E}^k_{alt}} \sum_{j=1}^{k} E_{\nu\pi}[T_j(\tau)] D(\nu_j, \nu'_j)$$

Stopping criteria motivated by (2): Stop at first

$$t \geq k : \quad Z_t > \beta_t(\delta)$$

where $\beta_t(\delta) \simeq \underbrace{2k(\ln t)}_{h(t) \text{ is slow wrt } t;\ Z_t \text{ grows linearly in } t} + \ln(1/\delta) + \underbrace{o\left(\ln t + \ln \frac{1}{\delta}\right)}_{\text{exact correction to be specified later.}}$

(2) Choosing $\pi = (\pi_t, t = 1, 2, 3, \ldots)$, the sampling strategy:

Since LB uses $\alpha_j(\nu) = \frac{E_{\nu\pi}[T_j(\tau)]}{E_{\nu\pi}[\tau]}$, we will

try to satisfy for any $t \geq k$: $T_j(t) \approx \alpha_j^*(\nu) \cdot t$.

The problem is that we do not know ν. However, we do have an approximation $\nu \approx \hat{\nu}(t) = (\hat{\mu}_1(t) \cdots \hat{\mu}_k(t))$

$\therefore$ Choose action $A_{t+1} = \underset{1 \leq j \leq k}{\arg\max} \left(t\, \hat{\alpha}_j^*(t) - T_j(t) \right)$

where $\hat{\alpha}_j^*(t) = \alpha^*(\hat{\nu}(t))$

i.e.,

$$\alpha^*(\hat{\nu}(t)) = \arg\sup_{\alpha \in \rho^{k-1}} \inf_{\nu' \in \mathcal{E}^k_{alt}(\hat{\nu}(t))} \sum_{j=1}^{k} (t\alpha_j) \frac{(\hat{\mu}_j(t) - \mu_j')^2}{2}$$

$\mathcal{E}^k_{alt}(\hat{\nu}(t)) = \{ (\mu_1' \mu_2' \cdots \mu_k') : \exists j = i^*(t) \text{ s.t. } \mu_j' > \mu'_{i^*(t)} \}$

As above, if there are more than one best arm, pick one of them to be the best (breaking ties arbitrarily).

(3) Ensure sufficient number of samples for each arm for sufficient concentrations, so that above steps of "replace expectations by sample-average" work!

At any time t s.t. $T_j(t) \leq \sqrt{t}$, for some $j \in \{1, 2, \cdots, k\}$, choose

$$A_{t+1} = \underset{1 \leq j \leq k}{\operatorname{argmin}} \; T_j(t).$$

④ Decision ψ : Stop when ① above is violated.

$$\psi = i^*(\nu(t)) = \underset{1 \leq j \leq k}{\operatorname{argmax}} \; \hat{\mu}_j(t)$$

Return:

$$\tau = t$$

Track-and-Stop Algorithm

(Garivier & Kaufmann 2016)

Input: δ, $\beta_t(\delta)$

Initial: Play each arm once (k rounds total).

Iteration: If $Z_t < \beta_t(\delta)$, then:

Case (i) $\operatorname{argmin}_{1 \le j \le k} T_j(t) \le \sqrt{t}$: $A_{t+1} = \operatorname{argmin}_j T_j(t)$

Case (ii) $\operatorname{argmin}_{1 \le j \le k} T_j(t) > \sqrt{t}$: $A_{t+1} = \operatorname{argmax}_j \left(t \hat{\alpha}_j^*(t) - T_j(t) \right)$

Observation: Reward sample $X_{A_{t+1}(t)}(t)$, update arm statistics

$t \to t+1$

Return: $\psi = \operatorname{argmax}_j \hat{\mu}_j(t)$, $\tau = t$

Choice of $\beta_t(\delta)$: We earlier discussed that $\beta_t(\delta)$ should scale as $2k \ln t + \ln(1/\delta) + o(\cdots)$.

More precisely, let $f(x) = e^{k-x} \cdot \left(\frac{x}{k}\right)^k$, and

$$\beta_t(\delta) = k \ln(t^2 + t) + f^{-1}(\delta)$$

$\hookrightarrow \approx (1 + o(1)) \ln(1/\delta)$

Lemma below $\Rightarrow$ Track-and-Stop is sound.

Lemma (33.7 in text): $\beta_t(\delta)$ as above, and

$$\tau = \min\{t \geq k : Z_t \geq \beta_t(\delta)\}.$$ Then

$$P\left(i^*(\hat{\nu}(\tau)) \neq i^*(\nu)\right) \leq \delta$$

$i^*(\hat{\nu}(\tau))$: estimated best arm.

$i^*(\nu)$: true best arm in environment ν

This also formally justifies intuition ① on the selection of the stopping time. To prove this, we need some concentration results.

Lemma on concentration (33.8 in text): $\{X_t, t=1,2,3,\ldots\}$ iid $N(\mu, 1)$ rvs. $\hat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$. Then:

$$P\left(\exists\, n \geq 1 \text{ s.t. } (\hat{\mu}_n - \mu)^2 \geq \frac{2}{n}\left(\ln(1/\delta) + \ln(n^2+n)\right)\right) \leq \delta.$$

Pf: Gaussian tail bound + Union bound. ∎

Prop. (33.9 in text): $g: \mathbb{N} \to \mathbb{R}$ an increasing function. Further, define $\{S_{i1}, S_{i2}, \dots\}$, $i = 1, 2, \dots, k$, sequences of independent r.v.s. s.t. for any $\delta \in (0,1)$, this sequence satisfies:

$$P\left(\exists\, r \in \mathbb{N} : S_{ir} \geq g(r) + \ln\left(\tfrac{1}{\delta}\right)\right) \leq \delta$$

Then, for any $x \geq 0$, we have:

$$P\left(\exists \{r_1, r_2 \dots r_k\} \in \mathbb{N}^k : \sum_{i=1}^{k} S_{i r_i} \geq k\, g\left(\sum_{i=1}^{k} r_i\right) + x\right)$$

$$\leq \left(\frac{x}{k}\right)^k e^{(k-x)}$$

Proof:

Use the fact: If $\{X_t, t \geq 1\}$ indep with $P(X_t \leq a) \leq a$ $\forall\, a \in [0,1]$, then, for any $\varepsilon > 0$,

$$P\left(\sum_{t=1}^{n} \ln\left(\frac{1}{X_t}\right) \geq \varepsilon\right) \leq \left(\frac{\varepsilon}{n}\right)^n e^{(n-\varepsilon)}$$

(Problem 5.16 in text).

Now, let $W_i = \max\left\{w \in [0,1] : S_{ir} < g(s) + \ln\left(\tfrac{1}{w}\right)\ \forall r = 1, 2, \dots\right\}$,

for $i = 1, 2, \cdots, K$. Then for any $\{r_1, r_2 \cdots, r_K\}$ positive integers, we have

$$\sum_{i=1}^{K} S_i r_i \leq \sum_{i=1}^{K} g(r_i) + \sum_{i=1}^{K} \ln\left(1/w_i\right)$$

$$\leq K g\left(\sum_{i=1}^{K} r_i\right) + \sum_{i=1}^{K} \ln\left(1/w_i\right)$$

trivially, as $g(\cdot)$ is increasing

Now, $\{W_i, i=1, 2, \cdots, K\}$ indep. and $P(W_i \leq a) = a$ $\forall a \in [0,1]$. Thus, use fact above ∎

Back to Lemma (33.7); pp. 23 in Notes above:

Need to show that: $P\left(i^*(\hat{\nu}(\tau)) \neq i^*(\nu)\right) \leq \delta$

First note that the Track-and-Stop algorithm will output a unique $i^*(\hat{\nu}(\tau))$. Specifically, if there exists two arms that have $\hat{\mu}_i(t) = \hat{\mu}_j(t)$ at some time $t \geq 1$, then $Z_t = 0$, and thus $0 = Z_t < \beta_t(\delta)$ ($\beta_t(\delta) > 0$). Thus, we will not stop at this time.

WLOG, assume $\Delta_1 = 0$ (i.e., arm 1 is the best arm),

Then, with $\mu_j' \equiv \mu_j'(\nu')$, (i.e., the mean of arm j in environment ν')

By the defn of (Z_t, τ), we have:

$$Z_\tau = \inf_{\nu' \in \mathcal{E}_{alt}^k(\hat{\nu}(\tau))} \sum_{j=1}^k T_j(\tau) \frac{(\hat{\mu}_j(\tau) - \mu_j')^2}{2} \geq \beta_\tau(\delta).$$

$$\Rightarrow \forall\, \nu' \in \mathcal{E}_{alt}^k(\hat{\nu}(\tau)), \quad \sum_{j=1}^k T_j(\tau) \frac{(\hat{\mu}_j(\tau) - \mu_j')^2}{2} \geq \beta_\tau(\delta).$$

$$\Rightarrow$$

$$\left\{\nu' \in \mathcal{E}_{alt}^k(\hat{\nu}(\tau))\right\} \subseteq \left\{\nu' \in \mathcal{E} : \frac{1}{2} \sum_{j=1}^k T_j(\tau)(\hat{\mu}_j(\tau) - \mu_j')^2 \geq \beta_\tau(\delta)\right\}$$

Next, observe that the error event:

$$\left\{1 \notin i^*(\hat{\nu}(\tau))\right\} = \left\{\nu \in \mathcal{E}_{alt}^k(\hat{\nu}(\tau))\right\}$$

(1: true best arm; ν: true environment)

Combining above:

(μ_j: true mean)

$$P\left(1 \notin i^*(\hat{\nu}(\tau))\right) \leq P\left(\frac{1}{2} \sum_{j=1}^k T_j(\tau)(\hat{\mu}_j(\tau) - \mu_j)^2 \geq \beta_\tau(\delta)\right)$$

Combining with the concentration results above, the lemma now follows. ▨

Theorem (33.6 in text): (π, τ, ψ) according to the Track-and-Stop policy. $\beta_t(\delta)$ chosen as above. Further $\mathcal{E} = \mathcal{E}_N^k(1)$, with $|i^*(\nu)| = 1$. (WLOG, $i^*(\nu) = 1$). Then:

(i) Track-and-Stop is sound.

(ii) $$\lim_{\delta \to 0} \frac{E_{\nu\pi}[\tau]}{\ln(1/\delta)} = c^*(\nu)$$ (asymp. optimal).

Proof:

(i) Soundness proved in Lemma 33.7 (above).

(ii) (Sketch of proof):

$$Z_t = \inf_{\nu \in \mathcal{E}_{alt}^k(\hat{\nu}(t))} \frac{1}{2} \sum_{j=1}^k T_j(t) \left(\hat{\mu}_j(t) - \mu_j'\right)^2$$

$\approx (\alpha_j^*(\nu) \cdot t)$, assuming that we have sufficient samples s.t. $\hat{\nu}(t) \approx \nu$

$$\approx t \inf_{\nu' \in \mathcal{E}_{alt}^k(\hat{\nu}(t))} \frac{1}{2} \sum_{j=1}^k \alpha_j^*(\nu) \left(\mu_j(\nu) - \mu_j'\right)^2$$

$\hat{\nu}(t)$: assume $\hat{\nu}(t) \approx \nu$

$\mu_j(\nu)$: assuming $\hat{\mu}_j(t) \approx \mu_j(\nu)$

$$= t \cdot \left(c^*(\nu)\right)^{-1}$$

$\therefore$ At $t = \tau$,

ignore as small wrt τ

$$\frac{\tau}{c^*(\delta)} \approx Z_\tau \geq \beta_t(\delta) \approx 2k \ln \tau + \ln\left(\tfrac{1}{\delta}\right) + o\left(\ln(\tau/\delta)\right)$$

$\approx$: from above intuition

$\geq$: stopping criterion

$o(\ln(\tau/\delta))$: ignore as small wrt anything else

i.e., as $\delta \to 0$, $\quad \frac{\tau}{c^*(\delta)} \approx \ln\left(\tfrac{1}{\delta}\right)$

Fixed Horizon Bandit: Best Arm Identification.

Setting: Similar as before, but stop at a fixed pre-determined (known) time and declare one of the arms as the best one.

Algorithm: Elimination $\equiv$ Successive Halving Algorithm.

Time horizon (n) divided into L rounds, where $L = \lceil \log_2 k \rceil$. In each round, eliminate half the arms.

Input: n, $L = \lceil \log_2 k \rceil$.

Initial: $A_1 = \{1, 2, \dots, k\}$.

Iteration: For each $\ell = 1, 2, \dots, L$

a. $T_\ell = \left\lfloor \frac{n}{L \cdot |A_\ell|} \right\rfloor$

b. Draw T_ℓ samples for every arm $j \in A_\ell$.

c. Compute $\hat{\mu}_j^\ell$ for each $j \in A_\ell$ using the T_ℓ samples from this round.

d. $A_{\ell+1} \subseteq A_\ell$: top $\left\lceil \frac{|A_\ell|}{2} \right\rceil$ arms

Termination: Return sole surviving arm.

Theorem: $\nu \in \mathcal{E}_{SG}^k(1)$, with $\mu_1 \geq \mu_2 \geq \dots \geq \mu_k$, and π the algorithm above. Then:

$$P_{\nu\pi}\left(\Delta_{A_{n+1}} > 0\right) \leq 3 \log_2 k \, e^{-n / (16 H_2(\mu) \log_2 k)}$$

i.e. finds a best arm

where $H_2(\mu) = \max_{\{i:\Delta_i>0\}} \frac{i}{\Delta_i^2}$.