What's new at Inria?
DELTA Project Meeting

Liège, April 30th, 2019

People

Permanent researcher:

- Michal Valko (Inria $\rightarrow$ DeepMind... but still in Delta)
- Emilie Kaufmann (CNRS)

PhD candidate:

- Omar Darwiche-Domingues (since October 2018)

Post-doc:

- Pierre Ménard (since February 2019)

Several collaborators

Agenda

Task 2.1, “Best Arm Identification Tools for Planning”

1. News (BAI) tools that can be useful for Planning
2. Keeping Non-Stationarity in Mind
3. Recent Work on Planning (and the Simulator)

Agenda

1. News (BAI) tools that can be useful for Planning
2. Keeping Non-Stationarity in Mind
3. Recent Work on Planning (and the Simulator)

The Power of Zipf Sampling

Abbasi-Yadkori, Bartlett, Gabillon, Malek and Valko. *Best of both worlds: Stochastic & adversarial best-arm identification*, COLT'18

For $t = 1, 2, \ldots$

- Sort and rank the arms by decreasing order of estimated cummulated gain $\hat{G}_k(t-1)$: Rank arm k as $\langle \tilde{k} \rangle_t$
- Select arm $A_t \in [K]$ at random such that

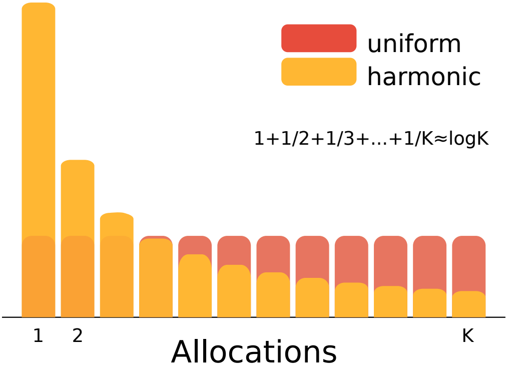
$$\mathbb{P}\left(A_t = k\right) = \frac{1}{\langle \tilde{k} \rangle_t \overline{\log}(K)}$$

Recommend, at any given round t,

$$J_t \triangleq \operatorname*{arg\,max}_{k \in \{1,\ldots,K\}} \hat{G}_k(t).$$

Figure: The P1 algorithm

The Power of Zipf Sampling

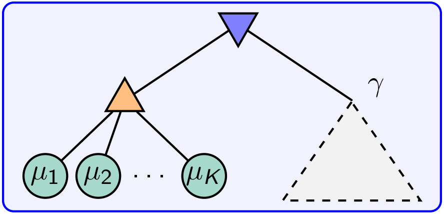


Variants of this simple allocation rule has already been used in different settings:

- black-box optimization (SequOOL, ALT'19)
- planning (PlatγPOOs, ICML' 19)

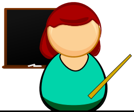
Some new insights from a toy Active Testing Problem

Kaufmann, Koolen and Garivier, *Sequential Test for the Lowest Mean: from Thompson to Murphy Sampling*, NeurIPS'18



Fix threshold γ.

$\mu^* := \min_i \mu_i \lessgtr \gamma$?



For $t = 1, \dots, \tau$

- pick an arm A_t
- observe $X_t \sim \mu_{A_t}$

After stopping, recommend $\hat{m} \in \{<, >\}$

Goal: controlled error $\mathbb{P}_{\boldsymbol{\mu}} \{\text{error}\} < \delta$ and small sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

Lower Bound and Oracle Allocation

Generic lower bound [Garivier et al. 16] shows *sample complexity* for *any* δ-correct algorithm is at least

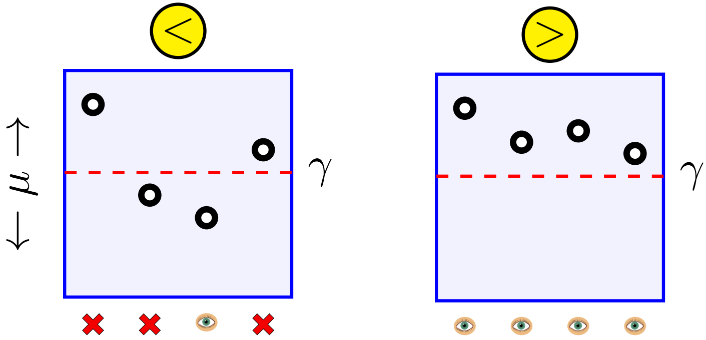
$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln\left(\tfrac{1}{\delta}\right).$$

For our problem the *characteristic time* and *oracle weights* are

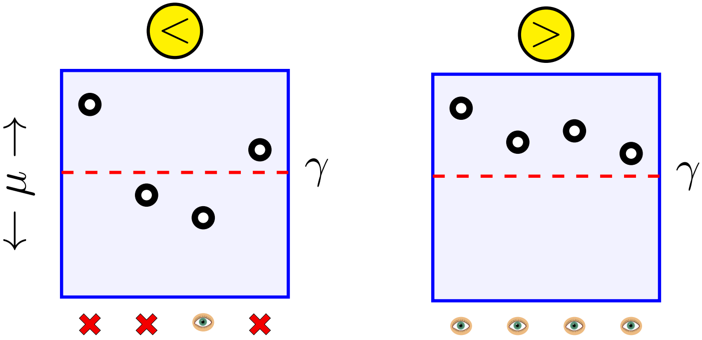
$$T^*(\boldsymbol{\mu}) \;=\; \begin{cases} \dfrac{1}{d(\mu^*, \gamma)} & \mu^* < \gamma, \\ \displaystyle\sum_a \frac{1}{d(\mu_a, \gamma)} & \mu^* > \gamma, \end{cases} \qquad w_a^*(\boldsymbol{\mu}) \;=\; \begin{cases} \mathbf{1}_{(a=a^*)} & \mu^* < \gamma, \\ \frac{\frac{1}{d(\mu_a,\gamma)}}{\sum_j \frac{1}{d(\mu_j,\gamma)}} & \mu^* > \gamma. \end{cases}$$

$w_a^*(\boldsymbol{\mu})$: fraction of selections of arm a under a strategy that would match the lower bound

Dichotomous Oracle Behaviour! Sampling Rule?



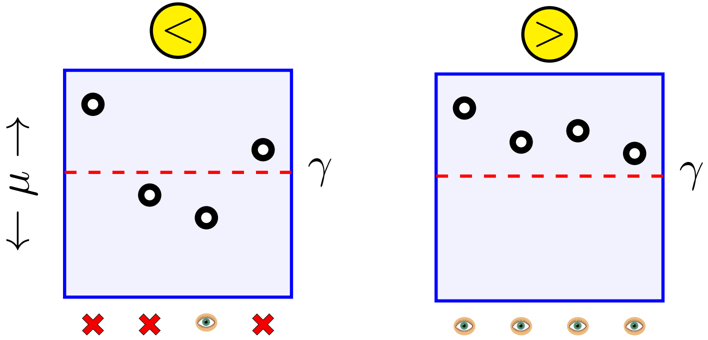
Dichotomous Oracle Behaviour! Sampling Rule?



Two different ideas to get those sampling profiles:

- **Thompson Sampling** (Π_{t-1} is posterior after $t-1$ rounds)
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}$, then play $A_t = \arg\min_a \ \theta_a$.
- **a Lower Confidence Bound algorithm**
  Play $A_t = \arg\min_a \ \mathrm{LCB}_a(t)$

A Solution: Murphy Sampling!



A more flexible idea:

- **Murphy Sampling** **condition on low minimum mean**
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}\left(\cdot \middle| \min_a \theta_a < \gamma\right)$, then play $A_t = \arg\min_a \ \theta_a$.

$\rightarrow$ converges to the optimal allocation in both cases!

Properties of Murphy Sampling [KKG, NeurIPS'18]

Theorem

Asymptotic optimality: $N_a(t)/t \to w_a^*(\boldsymbol{\mu})$ *for all* $\boldsymbol{\mu}$

Sampling rule	$<$	$>$
Thompson Sampling	✓	✗
Lower Confidence Bounds	✗	✓
Murphy Sampling	✓	✓

Lemma

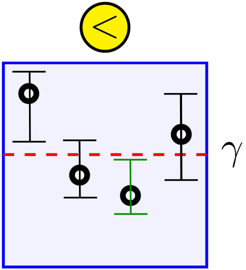
Any anytime sampling strategy $(A_t)_t$ *ensuring* $\frac{\boldsymbol{N}_t}{t} \to \boldsymbol{w}^*(\boldsymbol{\mu})$ *and good stopping rule* τ_δ *guarantee* $\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln \frac{1}{\delta}} \leq T^*(\boldsymbol{\mu})$.

$\to$ Murphy Sampling combined with a good stopping rule asymptotically attains the optimal sample complexity.

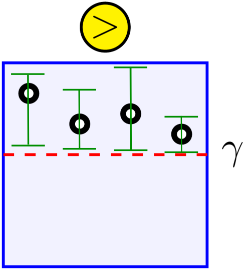
What is a "good stopping rule"?

Example: a stopping rule based on individual confidence bounds:
$\tau^{\text{Box}} := \min(\tau_<; \tau_>)$ where

$$\tau_< = \inf\{t \in \mathbb{N} : \exists a : \mathrm{UCB}_a(t) < \gamma\}$$
$$\tau_> = \inf\{t \in \mathbb{N} : \forall a, \mathrm{LCB}_a(t) > \gamma\}$$



$\tau = \tau_<$



$\tau = \tau_>$

What is a "good stopping rule"?

Example: a stopping rule based on individual confidence bounds:
$\tau^{\text{Box}} := \min(\tau_<; \tau_>)$ where

$$\tau_< = \inf\{t \in \mathbb{N} : \exists a : \mathrm{UCB}_a(t) < \gamma\}$$
$$\tau_> = \inf\{t \in \mathbb{N} : \forall a, \mathrm{LCB}_a(t) > \gamma\}$$

➜ enough to have the previous (asymptotic) results, but in practice we want to leverage the following:

Multiple low arms
identical or similar

$\Rightarrow$
- conclude $\mu^* < \gamma$ **faster**
- **tighter** confidence interval for μ^*

Improved Upper Confidence Bound on a Minimum

Given a subset $\mathcal{S} \subseteq \{1, \dots, K\}$, let

- $N_{\mathcal{S}}(t)$ the number of selections of an arm in $\mathcal{S}$
- $\hat{\mu}_{\mathcal{S}}(t)$ the **aggregated empirical mean**

Theorem

For any prior π, for an appropriate choice of threshold $\mathcal{T}$,

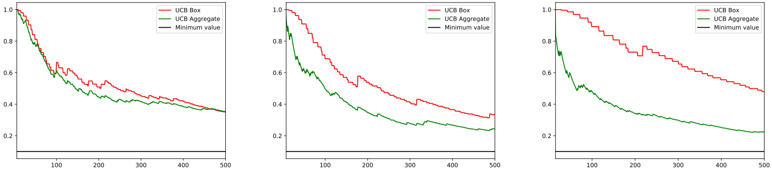
$$\mathrm{UCB}^{\pi}_{\min}(t) := \max\Big\{q : \exists \mathcal{S} \subseteq [K] : \left[N_{\mathcal{S}} d^{+}(\hat{\mu}_{\mathcal{S}}, q) - \ln\ln N_{\mathcal{S}}\right] \leq \mathcal{T}\left(\ln \frac{1}{\delta \pi(\mathcal{S})}\right)\Big\}$$

satisfies $\mathbb{P}\left(\forall t \in \mathbb{N}, \mathrm{UCB}^{\pi}_{\min}(t) > \mu^{*}\right) \geq 1 - \delta$.

Improved stopping rule:

$$\tau_{<} = \inf\{t \in \mathbb{N} : \mathrm{UCB}^{\pi}_{\min}(t) \leq \gamma\}$$

Improved Upper Confidence Bounds on a Minimum



UCB for minimum: Agg *dominates* Box with 1, 3 and 10 low arms.

(on can also get a larger LCB on the maximum mean)

Agenda

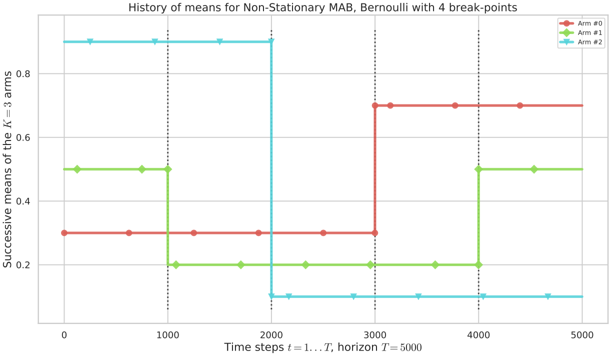
The Complexity of Rotting Bandits

Rotting bandits: each time an arm is played, its mean decreases (a specific form of non-stationnarity)

Seznec, Locatelli, Carpentier, Lazaric, Valko. *Rotting bandits are not harder than stochastic ones*, AISTATS'19 (oral presentation)

(the reason Michal is not here today...)

Piecewise-Stationary Model: one Example



nb of breakpoints: $\Upsilon_T = 4$

Positioning

(Quick) related work

- Existing guarantees for a variant of EXP3
  EXP3.S [Auer et al. 2002]
- Still, many attempts to adapt *stochastic bandit algorithms* to this problem: CUSUM-UCB [Liu et al, 2018], Monitored-UCB [Cao et al, 2019]
- Those attemps require the knowledge of

the number of breakpoints + a lower bound on the minimal magnitude of change

Positioning

(Quick) related work

- Existing guarantees for a variant of EXP3
  EXP3.S [Auer et al. 2002]
- Still, many attempts to adapt *stochastic bandit algorithms* to this problem: CUSUM-UCB [Liu et al, 2018], Monitored-UCB [Cao et al, 2019]
- Those attemps require the knowledge of

the number of breakpoints + a lower bound on the minimal magnitude of change

Our contributions:

- kl-UCB + un efficient adaptive sliding window
- no need to know anything about the size of a change

The Bernoulli GLRT

Context: Piecewise i.i.d. bandit with bounded rewards.

Key tool: an efficient change-point detector to detect a change in the mean of a bounded distribution, the Bernoulli-GLR

➜ given a stream of samples $(X_s) \in [0,1]$, detection occurs after n samples if

$$\sup_{s \in [1,n]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) \right] \geq \beta(n, \delta)$$

where $\hat{\mu}_{s:s'} = (\sum_{k=s}^{s'} X_s)/(s'-s+1)$ and

$$\mathrm{kl}(x,y) = x \ln(x/y) + (1-x) \ln((1-x)/(1-y)).$$

The Bernoulli GLRT

Context: Piecewise i.i.d. bandit with bounded rewards.

Key tool: an efficient change-point detector to detect a change in the mean of a bounded distribution, the Bernoulli-GLR

➜ given a stream of samples $(X_s) \in [0,1]$, detection occurs after n samples if

$$\sup_{s\in[1,n]} \left[s \times \mathrm{kl}\left(\hat{\mu}_{1:s}, \hat{\mu}_{1:n}\right) + (n-s) \times \mathrm{kl}\left(\hat{\mu}_{s+1:n}, \hat{\mu}_{1:n}\right) \right] \geq \beta(n,\delta)$$

Lemma (false alarm probability)

For $\beta(n,\delta) \simeq \ln(3n\sqrt{n}/\delta)$ the probability that a detection occurs on a i.i.d. stream is at most δ.

kl-UCB meets the GLRT

Parameters: $\alpha \in (0,1)$, $\delta > 0$.
Arm selection: at round t,

- if $\alpha > 0$ and $t \mod \lfloor K/\alpha \rfloor \in \{1, \dots, K\}$,

$$(\textit{forced exploration}) \quad A_t \leftarrow t \mod \lfloor K/\alpha \rfloor$$

- else, select

$$(\textit{kl-UCB}) \qquad A_t \leftarrow \arg\max_a \mathrm{UCB}_a(t)$$

$\tau_a(t)$	:	instant of the last restart
$n_a(t)$	:	number of selection of arm a since the last restart
$\hat{\mu}_a(t)$	:	empirical mean of samples from arm a since last restart

$$\mathrm{UCB}_a(t) := \max\left\{q \in [0,1] : n_a(t) \times \mathrm{kl}\left(\hat{\mu}_a(t), q\right) \leq f(t - \tau_a(t))\right\}.$$

Restarts: Local or Global after a change is detected by the Bernoulli-GLRT on the mean of the selected arm

Results

- a unified analysis of Local and Global changes
- a tuning of the algorithm that ensures $O(\Upsilon_T \sqrt{T})$ when Υ_T is unknown and $O(\sqrt{\Upsilon_T T})$ regret if Υ_T is known
- good practical performance !

work in progress with Lilian Besson (CentraleSupélec Rennes) and Odalric Maillard (Inria)

Agenda

1. News (BAI) tools that can be useful for Planning
2. Keeping Non-Stationarity in Mind
3. Recent Work on Planning (and the Simulator)

Planning in Regularized MDPs and Games

- Planning problem: given a generative model, estimate the value function at a state s;
- K actions, state space of any cardinality;
- We study value functions of entropy regularized MDPs and games.
- Example: Bellman equations for MDPs with entropy regularization

$$V(s) = \max_{\pi(\cdot|s) \in \mathcal{P}(\mathcal{A})} \mathbb{E}\left[r(s,a) + \gamma V(z)\right] + \lambda \underbrace{\mathcal{H}(\pi(\cdot|s))}_{\text{entropy}} \tag{1}$$

$$= \lambda \log \sum_{a \in \mathcal{A}} \exp\left(\frac{1}{\lambda}\mathbb{E}\left[r(s,a) + \gamma V(z)\right]\right), \; z \sim P(\cdot|s,a) \tag{2}$$

General case:

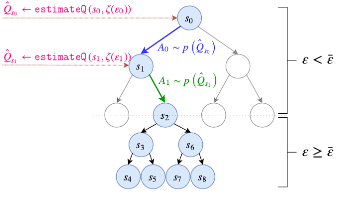
$$V(s) = F_s(Q_s), \quad \text{with} \quad Q_s(a) = \mathbb{E}\left[r(s,a) + \gamma V(z)\right], \; z \sim P(\cdot|s,a) \qquad (3)$$

- $F_s = \max$: Bellman equations for MDPs;
- $F_s = \max$ or min according to the player: value function for turn-based two-player game (discounted).
- Replace max and min by smooth approximations with LogSumExp = entropy regularization on the policy.
- Main assumptions:
  - (smoothness) $\left|F_s(x) - F_s(x_0) - (x - x_0)^T \nabla F_s(x_0)\right| \leq L||x - x_0||_2^2$;
  - F_s is 1-Lipschitz, nonnegative gradient.

Our algorithm

Algorithm 1 `sampleV`

Input: $(s, \varepsilon) \in \mathcal{S} \times \mathbb{R}_+$
if $\varepsilon \geq 1/(1-\gamma)$ **then**
  Output: 0
else if $\varepsilon \geq \bar{\varepsilon}$ **then**
  $\widehat{Q}_s \leftarrow$ `estimateQ`(s, ε)
  Output: $F_s\left(\widehat{Q}_s\right)$
else if $\varepsilon < \bar{\varepsilon}$ **then**
  $\widehat{Q}_s \leftarrow$ `estimateQ`$(s, \sqrt{\bar{\varepsilon}\varepsilon})$
  $A \leftarrow$ action drawn from $\frac{\nabla F_s(\widehat{Q}_s)}{\left\|\nabla F_s(\widehat{Q}_s)\right\|_1}$
  $(R, Z) \leftarrow$ `oracle`(s, A)
  $\widehat{V} \leftarrow$ `sampleV`$(Z, \varepsilon/\sqrt{\gamma})$
end if
Output: $F_s\left(\widehat{Q}_s\right) - \widehat{Q}_s^{\intercal} \nabla F_s\left(\widehat{Q}_s\right) + (R + \gamma \widehat{V}) \left\|\nabla F_s\left(\widehat{Q}_s\right)\right\|_1$



`SmoothCruiser`$(s_0, \varepsilon_0, \delta)$

Algorithm 2 `estimateQ`

Input: (s, ε)
// Compute the value of N
if $\varepsilon \geq 1/(1-\gamma)$ **then**
  Output: $(0, \ldots, 0)$
else if $\varepsilon \geq \bar{\varepsilon}$ **then**
  $N \leftarrow \frac{2}{(1-\gamma)^4(1-\sqrt{\gamma})^2} \frac{\log(2K/\delta)}{\varepsilon^2}$
else if $\varepsilon < \bar{\varepsilon}$ **then**
  $C \leftarrow \frac{1}{1-\gamma}\left(4\bar{\varepsilon} + \frac{4}{1-\gamma} + 1\right)$
  $N \leftarrow \frac{C^2}{2(1-\sqrt{\gamma})^2} \frac{\log(2K/\delta)}{\varepsilon^2}$
end if
// Average to estimate Q function
for $a \in \mathcal{A}$ **do**
  $q_i \leftarrow 0$ for $i \in 1, \ldots, N$
  for $i \in 1, \ldots, N$ **do**
    $(R, Z) \leftarrow$ `oracle`(s, a).
    $\widehat{V} \leftarrow$ `sampleV` $\left(Z, \varepsilon/\sqrt{\gamma}\right)$
    $q_i \leftarrow R + \gamma \widehat{V}$
  end for
  $\widehat{Q}_s(a) \leftarrow$ **mean** $(q_1, \ldots, q_N)$
end for
Output: $\widehat{Q}_s$

Algorithm 3 `SmoothCruiser`

Input: $(s, \varepsilon, \delta) \in \mathcal{S} \times \mathbb{R}_+ \times \mathbb{R}_+$
$\bar{\varepsilon} \leftarrow (1 - \sqrt{\gamma})/KL$
Set δ and $\bar{\varepsilon}$ as a global parameters
$\widehat{Q}_s \leftarrow$ `estimateQ`(s, ε)
Output: $F_s\left(\widehat{Q}_s\right)$

Guarantees

Theorem

Let $n(\epsilon, \delta)$ *be the number of calls to the generative model (oracle) before the algorithm terminates. For any state s and* $\epsilon, \delta > 0$,

$$n(\epsilon, \delta) \leq \frac{c_1}{\epsilon^4} \log\left(\frac{c_2}{\delta}\right) \left[c_3 \log\left(\frac{c_4}{\epsilon}\right)\right]^{\log_2\left(c_5\left(\log\left(\frac{c_2}{\delta}\right)\right)\right)} = \mathcal{O}\left(\frac{1}{\epsilon^{4+c}}\right), \ \forall c > 0$$

where c_1, c_2, c_3, c_4 *and* c_5 *are constants that depend only on* K, L *and* γ.

Theorem

For any state s and $\delta, \epsilon > 0$,

$$\mathbb{P}\left[\left|\hat{V}(s) - V(s)\right| > \epsilon\right] \leq \delta n(\epsilon, \delta).$$