Bandits for large scale problems

Alexandra Carpentier
Universität Potsdam, supported by DFG grant MuSyAD
(CA 1488/1-1)

Based on joint works with : **Rémi Munos, Michal Valko**

January 12, 2016

Bandits for large scale problems

In the classical bandit setting, it is usually assumed that the number of actions A is smaller than the horizon n, i.e.

$$A \leq n,$$

so that each action can be sampled at least once.

Here large scale problems are problems where $A \gg n$.

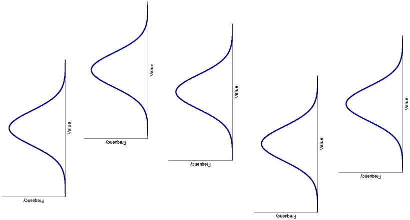
Outline

Outline

Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

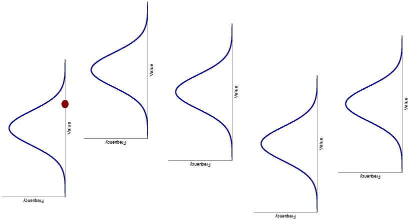
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

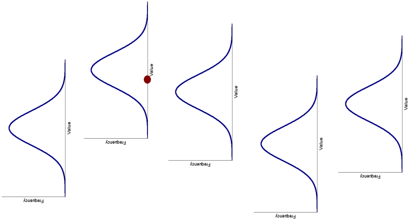
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

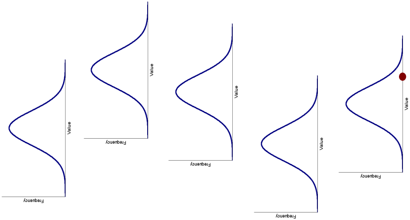
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

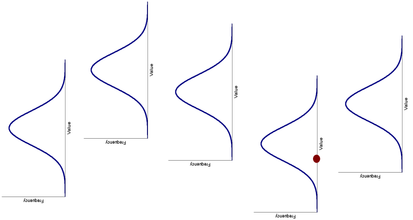
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

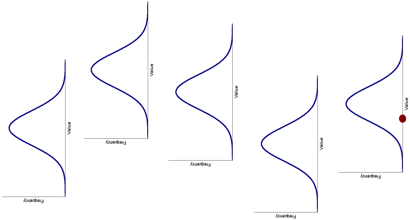
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

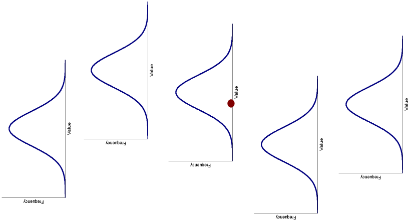
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

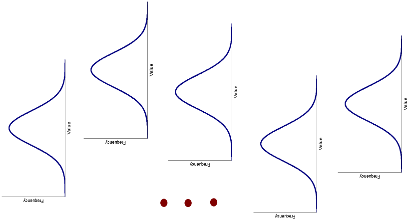
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

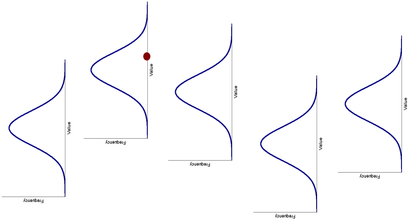
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

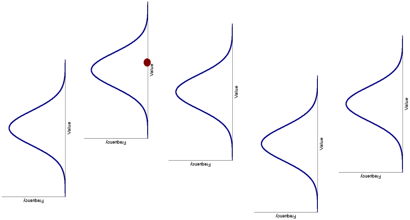
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

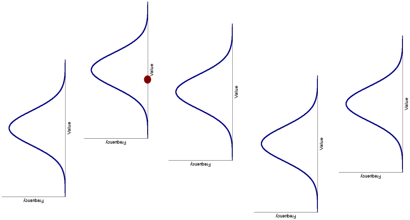
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

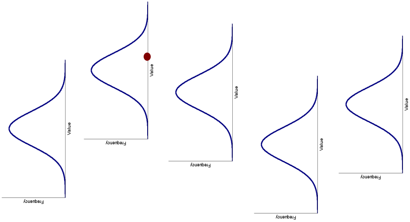
- Distributions $(\nu_a)_{a \le A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

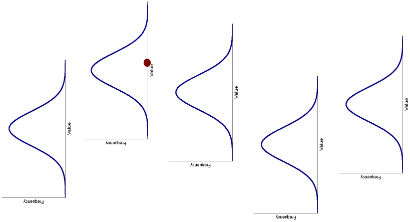
- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$



Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$

Objective of allocation when e.g. maximizing $\sum_t X_t$:

- Estimate all means μ_a of distributions (exploration)
- So that one finds the one with highest mean μ^* and samples it (exploitation)

Because of the noise to the samples, there is this exploration/exploitation trade-off.

Stochastic bandit setting

Resource allocation in face of uncertainty See [Thompson (1933)], [Robbins (1952)], [Gittins (1979)], etc.

- Distributions $(\nu_a)_{a \leq A}$ with *unknown* characteristics
- Limited sampling resources n
- At each time t, choose a_t and collect $X_t \sim \nu_{a_t}$
- Some objective, e.g. maximize $\sum_t X_t$

Popular solution to this trade-off is to sample the arm that maximizes an UCB [Auer et.al.(2002)] :

$$B_{a,t} = \hat{\mu}_{a,t} + c\sqrt{\frac{\log(n)}{T_{a,t}}}.$$

Theorem

The exp. regret is bounded as

$$\mathbb{E}R_n = n\mu^* - \mathbb{E}\sum_t X_t \leq c\sqrt{nA\log(n)}.$$

Stochastic bandit setting

Main question in this talk is on the *scale* of the problem.

Large scale aim

$$A \gg n.$$

Stochastic bandit setting

Main question in this talk is on the *scale* of the problem.

Large scale aim

$$A \gg n.$$

Possible alternative objectives :

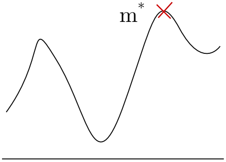
- Noisy optimisation Bubeck et al., 2010, Kaufman et al., 2012, Gabillon et al., 2012, Valko et al., 2013.
- Uniform functional estimation Antos et al., 2010, C et al., 2012, C et al., 2013.
- Stratified Monte-Carlo integration Grover et al., 2010, C et al., 2012, 2013, 2014.
- Extreme value detection Smith et al, 2009, C and Valko, 2014.

Noisy optimisation [Kleinberg et. al, 2008, Bubeck et al., 2010, Kaufman et al., 2012, Gabillon et al., 2012]

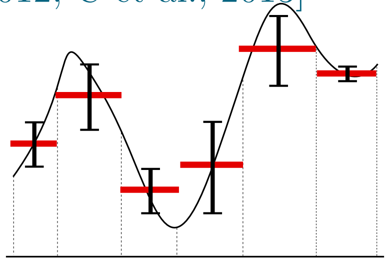
In the cumulative bandit setting, the objective is

$$\max \sum_t X_t.$$

A useful variant is the *pure exploration* variant of this setting where the aim is to return at the end of the budget $\hat{k}_n$ such that $\mu_{\hat{k}_n}$ is as large as possible (as close as possible to the optimal value μ^*). This is *noisy optimisation* in the bandit setting.



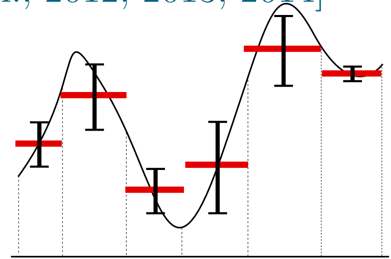
Adaptive stratified functional estimation [Antos et al., 2010, C et al., 2012, C et al., 2013]



Each stratum has measure w_k and sampling randomly in it results in a sample $X \sim \nu_k(\mu_k, \sigma_k^2)$.
Objective : Sample optimally in the strata to estimate the integral μ of the function and minimize

$$\max_k \mathbb{E}(\hat{\mu}_k - \mu_k)^2 = \max_k \frac{\sigma_k^2}{T_k}.$$

Adaptive stratified Monte-Carlo integration [Grover et al., 2010, C et al., 2012, 2013, 2014]



Each stratum has measure w_k and sampling randomly in it results in a sample $X \sim \nu_k(\mu_k, \sigma_k^2)$.
Objective : Sample optimally in the strata to estimate the integral μ of the function and minimize

$$\mathbb{E}(\hat{\mu}_n - \mu)^2 = \sum_k \frac{w_k^2 \sigma_k^2}{T_k}.$$

Outline

The "large scale" situation

Two main lines of work in order to solve this problem :

1. **Topological assumptions on the distributions :** There is a topology on the distributions so that information on a distribution provides information on other options as well. Examples :
   1.1 Linear topology.
   1.2 Smooth topology.
2. **No Topological assumptions on the distributions :** There is no topology on the options. Can represent also smooth topology in high dimension.

Linear topology : setting [Auer, 2002]

Problem : The set of arms $\mathcal{A}$ is a subset of $\mathbb{R}^D$, and $\alpha^* \in \mathbb{R}^D$ is an unknown parameter. At each time step t,

- Select $a_t \in \mathcal{A}$,
- Observe $X_t = \langle a_t, \alpha^* \rangle + \eta_t$, where $\mathbb{E}[\eta_t | a_t] = 0$.

Let $a^* = \arg\max_{a \in \mathcal{A}} \langle a, \alpha^* \rangle$ be the best arm in $\mathcal{A}$.
Define the regret:

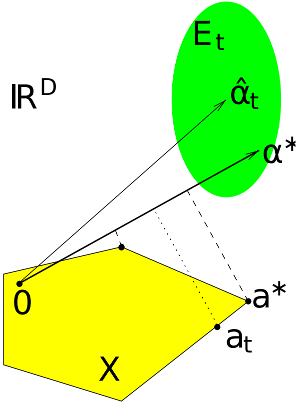
$$\mathbb{E}R_n = n\langle a^*, \alpha^* \rangle - \mathbb{E}\sum_{t=1}^{n} X_t.$$

No need to estimate the mean-reward of all arms, estimating α^* is enough [Auer, 2002], [Dani, Hayes, Kakade, 2008], [Abbasi-Yadkori, 2009], [Rusmevichientong, Tsitsiklis, 2010], [Filippi, Cappé, Garivier, Szepesvári, 2010].

Linear topology : UCB-based (ConfidenceBall) algorithm

Idea: Build a high probability confidence set E_t s.t. $\alpha^* \in E_t$ w.h.p. and play the arm $a \in \mathcal{A}$ that maximizes

$$B_{a,t} = \max_{\alpha \in E_t} \langle a, \alpha \rangle.$$



Linear Topology : Regret analysis and extensions

Theorem ((Dani, Hayes, Kakade, 2008, Rusmevichientong, Tsitsiklis, 2010))

The expected regret of ConfidenceBall is bounded as

$$\mathbb{E}R_n \leq D\sqrt{n}(\log n)^{3/2}$$

Possible extensions

- **Generalized Linear models** [Filippi, Cappé, Garivier, Szepesvári, 2010]..
- **Sparse linear bandits in high dimension** [C and Munos, 2012].

Extension to high dimensional and sparse linear bandits [C and Munos, 2012]

Linear bandit algorithm work if $D \ll n$. But what if $D \geq n$? In general nothing is possible but under the assumption that α^* is k-sparse and that $\mathcal{A}$ is the unit-ball, a solution is to first explore the space at random until the support of the signal is detected (CS phase) approximately, and then run ConfidenceBall on the right support (SL-UCB).

Theorem (C and Munos, 2012)

The expected regret of SL-UCB is bounded as

$$\mathbb{E}R_n \leq k\sqrt{n}(\log D)^{3/2}$$

Smooth topology : setting [Kleinberg et.al., 2008]

Problem: Let $f : \mathcal{A} \to \mathbb{R}$, assumed to be Lipschitz: $|f(x) - f(y)| \leq \ell(x, y)$.

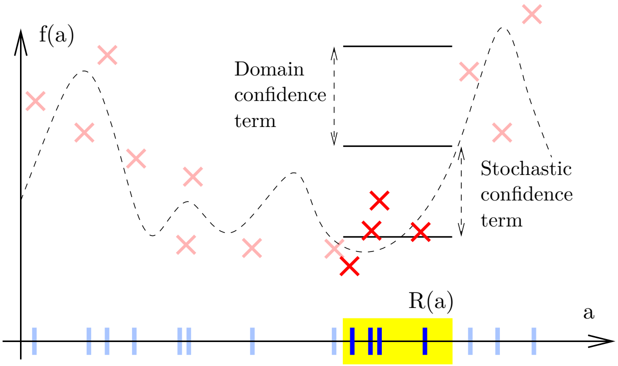
- At each time step t, select $a_t \in \mathcal{A}$
- Observe $X_t = f(a_t) + \eta_t$

Define the cumulative regret

$$R_n = nf^* - \sum_{t=1}^{n} X_t,$$

where $f^* = \sup_{a \in \mathcal{A}} f(a)$
Continuous stochastic optimization as a bandit problem [Kleinberg et.al., 2008, Srinivas et.al., 2009, Grünewälder et.al., 2010, Krause et.al.,2011, Bubeck et.al., 2010, Valko et.al.,2013].

Smooth topology : UCB-based (HOO) algorithm



Idea : Choose a small region $R(a)$ around a and sample the arm that maximizes

$$B_{a,t} = \hat{\mu}_{R(a),t} + SCT(a) + DCT(a).$$

Smooth Topology : Regret analysis

Theorem ((Kleinberg et.al., 2008, Bubeck et.al., 2010))

Let d be the **near-optimality dimension** *of f in $\mathcal{A}$: i.e. such that the set of ϵ-optimal actions*

$$X_\epsilon = \{x \in \mathcal{A}, f(x) \geq f^* - \epsilon\}$$

can be covered by $O(\epsilon^{-d})$ balls of radius ϵ.
The expected regret of HOO is bounded as

$$\mathbb{E}R_n \leq D n^{\frac{d+1}{d+2}}.$$

Extensions

- Unknown smoothness [Munos, 2013, Bull, 2014, Valko et al, 2015].
- Simple regret [Valko et.al., 2013].
- Continuous MC integration [C and Munos, 2013 a)b), 2014], [Pietquin et al., 2013].
- Uniform functional estimation [C and Maillard, 2013, Bull, 2013].

Continuous MC integration [C and Munos, 2014]

Assume that we want to integrate the function f and we can sample it n times and get at time t if sampling in x_t

$$y_t = f(x_t) + s(t)\eta_t,$$

where $\mathbb{V}(\eta_t) = 1$. The oracle optimal sampling stratgy has risk $\frac{\left(\int_{\mathcal{X}} s(x)dx\right)^2}{n}$.

Theorem

Assume that $|f(x) - f(y)| \leq \ell(x, y) = L\|x - y\|^\alpha$ *and* s *also* α*-Hölder and* $\mathcal{A} = [0, 1]^D$*. Then algorithm MC-ULCB outputing* $\hat{\mu}_n$ *estimating* $\int f$ *satisfies*

$$\mathbb{E}(\hat{\mu}_n - \int f)^2 - \frac{\left(\int s(x)dx\right)^2}{n} \leq CD^{\frac{2\alpha}{3d} + \frac{1}{2}}\sqrt{\log(n)}n^{-\frac{d+4\alpha}{d+3\alpha}}.$$

No topology : setting [Berry, Chen, Zame, Heath, Shepp, 1997]

Problem: Solve the stochastic bandit problem with $A \gg n$ (potentially $A = \infty$).

- At each time step t, select $a_t \in \mathcal{A}$
- Observe $X_t \sim \nu_{a_t}$

Define the cumulative regret

$$R_n = n\mu^* - \sum_{t=1}^{n} X_t,$$

where $\mu^* = \sup_{a \in \mathcal{A}} \mu_k$
Standard strategies do not apply when $A \gg n$ - need to sub-sample [Banks, Sundaram, 1992], [Berry, Chen, Zame, Heath, Shepp, 1997], [Wang, Audibert, Munos, 2008], [Bonald and Proutiere, 2008], [C and Valko, 2015].

No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

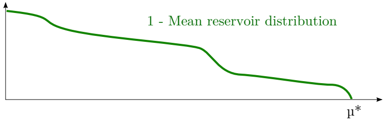
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 0$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

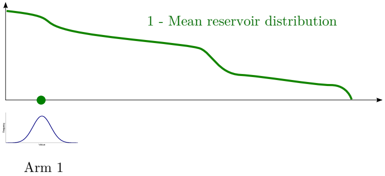
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 1$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

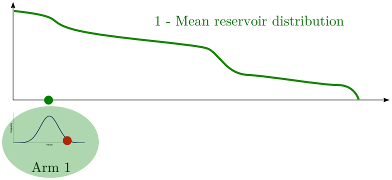
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 1$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

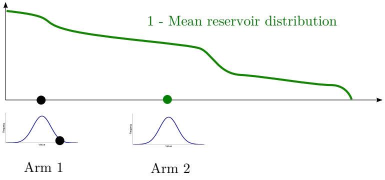
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 2$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

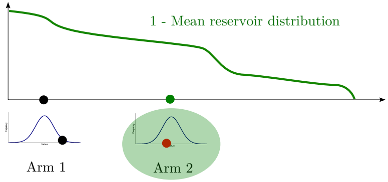
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 2$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

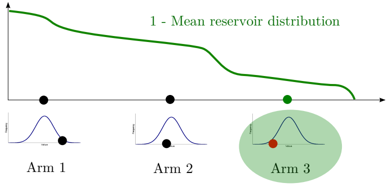
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 3$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

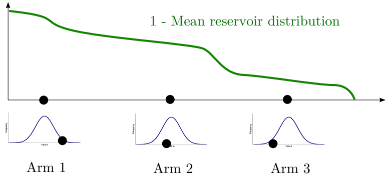
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 3$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

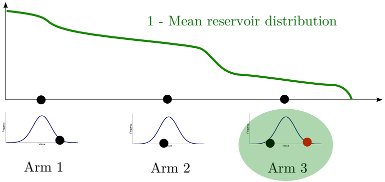
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 4$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

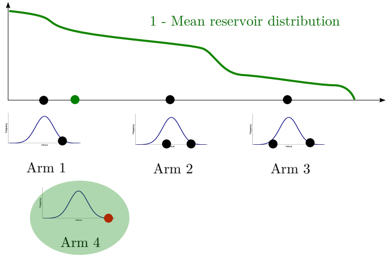
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 5$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

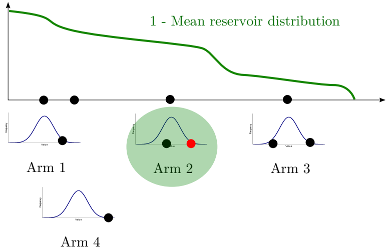
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 6$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

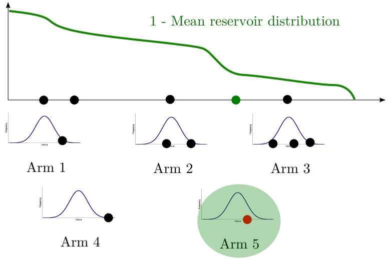
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 7$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

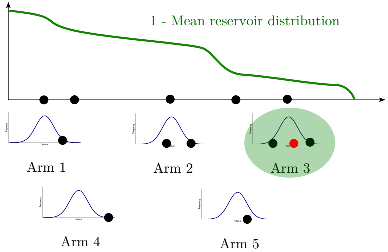
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 8$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

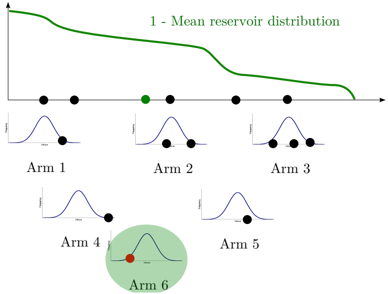
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time $t = 9$:



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

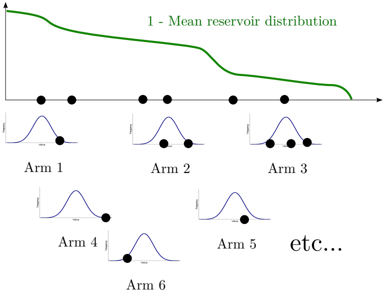
At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

Objective : Maximize $\sum_t X_t$.

At time t... :



No topology setting

- Arm reservoir distr. and an associated mean reservoir distr. F
- Limited sampling resources n, and $K_0 = 0$ observed arms

At time $t \leq n$ one can either

- set $K_t = K_{t-1} + 1$ and sample a new arm ν_{K_t} from the reservoir distr. with mean $\mu_{K_t} \sim F$, and set $I_t = K_t$,
- or choose an arm I_t among the K_{t-1} observed arms $\{\nu_k\}_{k \leq K_{t-1}}$,

and then collect $X_t \sim \nu_{k_t}$

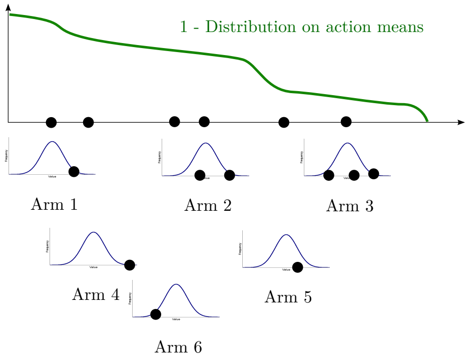
Objective : Maximize $\sum_t X_t$.

Double exploration and exploitation dilemma here : Allocation both to (i) learn the characteristics of the arm reservoir distr. (*meta-exploration*) and (ii) learn the characteristics of the arms (*exploitation*) and (iii) to maximize the sum of rewards (exploitation).

Main questions

How many arms should be sampled from the arm reservoir distribution? How aggressively should these arms be explored? What should be left for exploitation?

No topology : UCB-based (UCB-AIR) algorithm



Idea : Sub-sample the actions uniformly at random and adapt the number of actions to the proportion of sub-optimal actions.

No Topology : Regret analysis

Algorithm UCB-AIR : sub-sample $K_n \approx n^{\min(\beta/2, \beta/(\beta+1))}$ arms and sample the arm that maximize an UCB.

Theorem ((Wang, Audibert, Munos, 2008))

Assume that $\exists \beta > 0$ *such that*

$$\mathcal{P}(\mu(new\ arm) > \mu^* - \epsilon) \approx C\epsilon^{\beta}.$$

Then the expected regret of UCB-AIR is bounded as

$$\mathbb{E}R_n \leq C \max\left(\sqrt{n}, n^{\frac{\beta}{1+\beta}}\right).$$

Extensions : optimisation [C and Valko, 2015].

No topology and optimisation [C and Valko, 2015]

Problem: Return an arm $\hat{k}_n$ such that $\mu_{\hat{k}_n}$ is as large as possible.
Algorithm SiRI : sub-sample $K_n \approx n^{\min(\beta,2)/2}$ arms and sample the arm that maximize an UCB.

Theorem (C and Valko, 2015)

For SiRI we have up to $\log(n)$ *factors*

$$\mathbb{E}(\mu^* - \mu_{\hat{k}_n}) \leq \Big(\max\big(n^{-1/2}, n^{-\frac{1}{\beta}}\big)\Big).$$

Conclusion

Depending on the assumptions, many possible strategies.
Importance of :

- Minimal model assumptions
- Computational efficiency and simplicity
- Minimal calibration and versatility

Challenges :

- Good context integration
- Right assumptions
- Estimation of the regret of the srategies