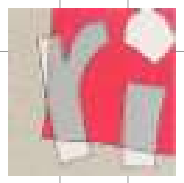


Exploration vs Exploitation challenge

Exploration, Exploitation, Evolution

Cédric Hartland, Michele Sebag

Université Paris Sud

The problem

A visitor comes to a website. The site provides options to the visitor who accepts or refuses.

- Representation of the visitor :
  - $p_i(t)$ =1 iff the visitor likes the option *i*
- p_{is} are independent and change along time

Goal : build $\hat{p}_i(t)$:
decision process : $t \rightarrow i = Max(\hat{p}_i(t))$

Applications : Mind reading machine (Shannon 1949)
pleasing a web site visitor

Our hidden motivation

Evolutionary computation (find Argmax(f: $\Omega \rightarrow R$)) :

- Solution population from the search space $\{X_1,...X_p\} \subset \Omega$
- Stochastic modifications on the population
  - Exploration : ~ random walk
  - Exploitation : ~ hill climbing, gradient

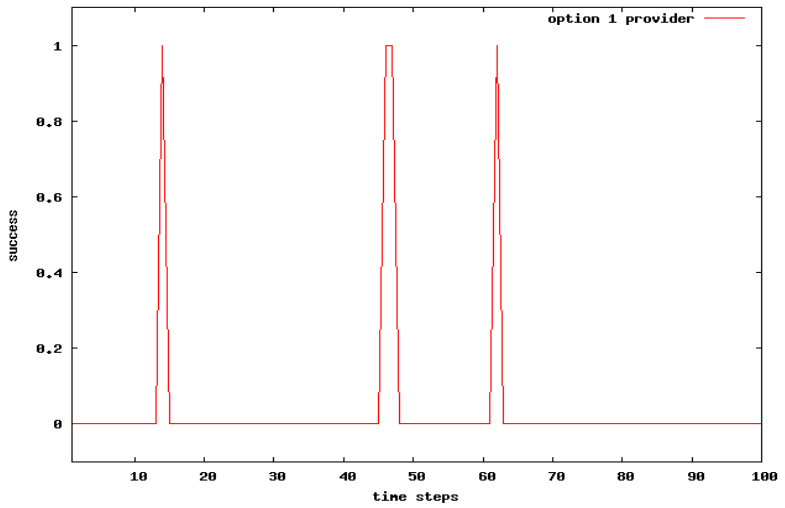
Problem : trade-off exploration exploitation (binary option)

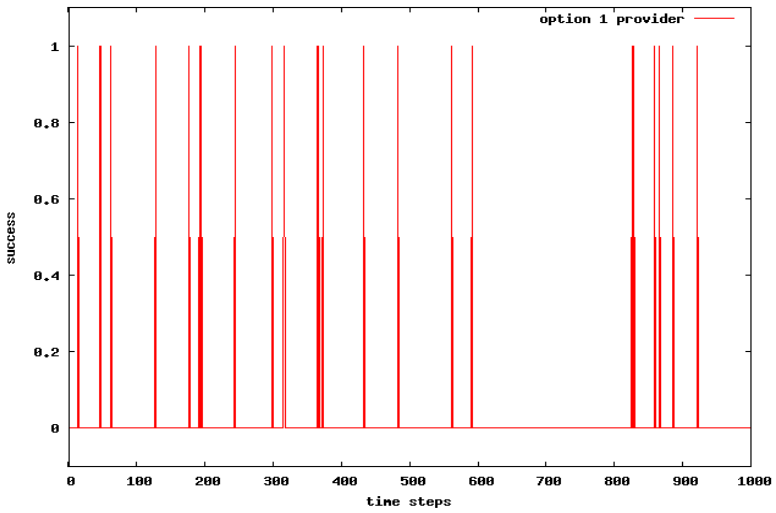
Problem formalisation : Holland 1975 : Multi-armed bandit simulation

BUT : dynamic reward

First step

- Preliminary exploration: experiments with naive models
  - Stick on one option
  - Switch with fixed frequencies
- Partial conclusions : no periodicity (obviously)
- Next step: Collect data to create a good model.

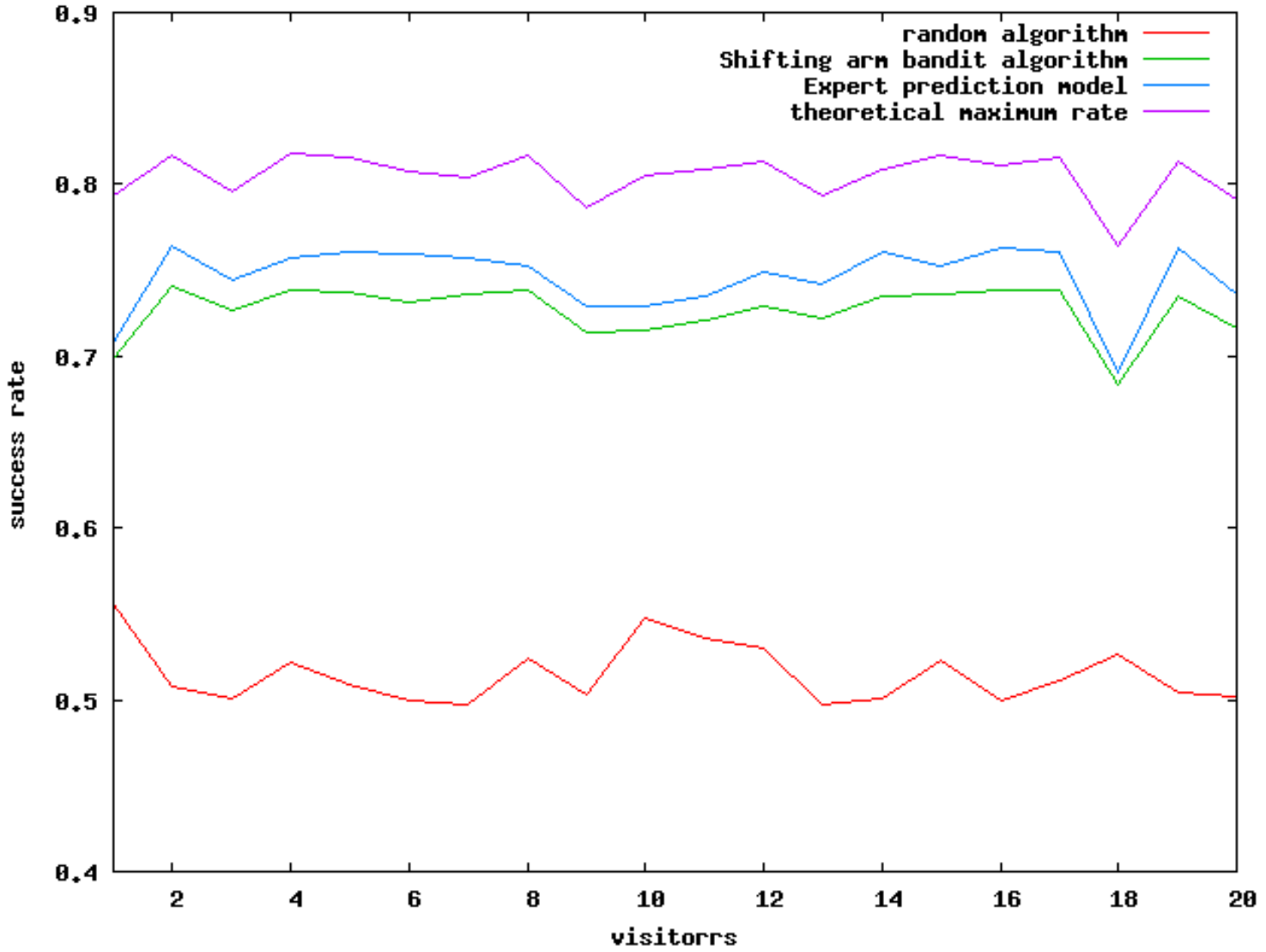




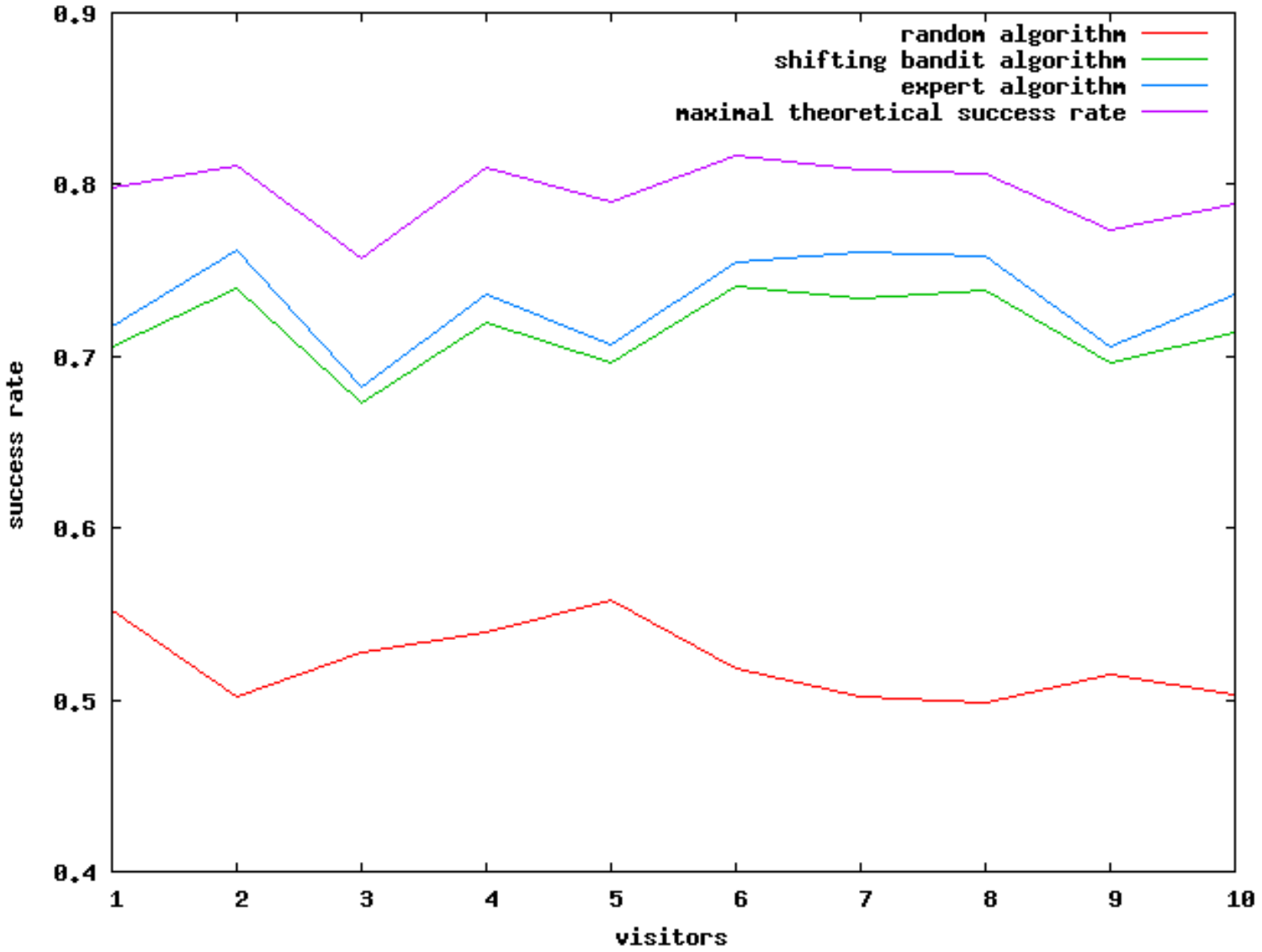
Mixture model

- Structure of expert i:
  - Time window on the past trials : 1..T=60
  - Average preference p_i in [0,1] → decision d_i
  - Weight w_i (confidence)
- Decision rule : weighted vote of the experts : result d
- Update rule : (over the time window)
  - If $d_i \neq d$: nothing [fading would be appropriate]
  - If $d_i = d$:
    - if failure then w_i /= W_f
    - if *success* then w_i *=W_s

Results


random algorithm
Shifting arm bandit algorithm
Expert prediction model
theoretical maximum rate
success rate
visitorrs
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
12
14
16
18
20

Results


random algorithm
shifting bandit algorithm
expert algorithm
maximal theoretical success rate
success rate
visitors
0.9
0.8
0.7
0.6
0.5
0.4
1
2
3
4
5
6
7
8
9
10

Results

Success rate + average regret on the testing set :

Algorithm	Success rate	Average Regret
Random provider	0.51	27031.8
Shifting-bandit provider	0.715	8020.3
Expert voting	0.732	6318.6

average theoretical maximal success rate : 0.796

Analysis of the results

Performance decreases if we reduce/increase the window/number of the experts

- Periodicity varies in this range ?
- Remark: At each time steps, for most experts p_i close to ½, some experts take the lead.

Limitations:

Fading

Learning on expert patterns

Conclusion - Perspectives

Online learning: weighted voting of experts is a good start

Next steps:

- Additive vs Multiplicative reward
- Recompute p_i with less weight on distant time steps.
- Varying reward factors
- Back to evolutionary computation: gain is continuous

Grazie