RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ONERA
THE FRENCH AEROSPACE LAB

Apprentissage et Planification sous Risque de Ruine

Filipo S. PEROTTO (DTIS / SYD)

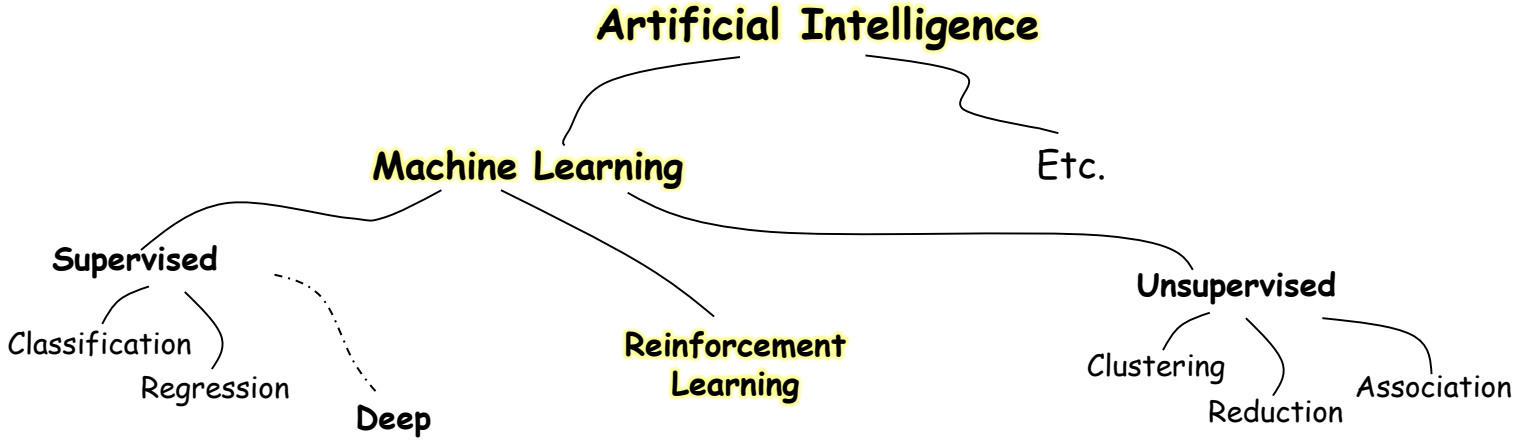
13 / 1 / 2022

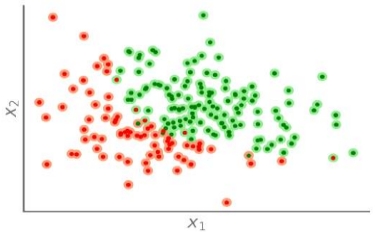
Dans cette présentation

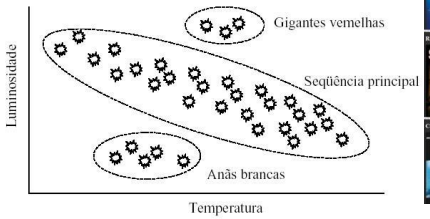
- **Introduction**
  - Panorama du **Machine Learning**
  - **Processus de Décision Markovien**
  - De l'**Apprentissage par Renforcement** à la **Ruine du Joueur**

- **Problèmes de Survie** (*budget safety*)
  - Perotto, F.S. et al. (2019). *Open Problem: Risk of Ruin in Multiarmed Bandits*. COLT 2019.
  - Perotto, F.S. et al. (2021). *Gambler Bandits and the Regret of Being Ruined*. AAMAS 2021.
  - Perotto, F.S. et al. (2021). *Deciding when to quit the gambler's ruin game with unknown probabilities*. IJAR. 137.
  - Perotto, F.S. et al. (2021) *A3R (Apprentissage par Renforcement et Risque de Ruine)*. Projet RG.

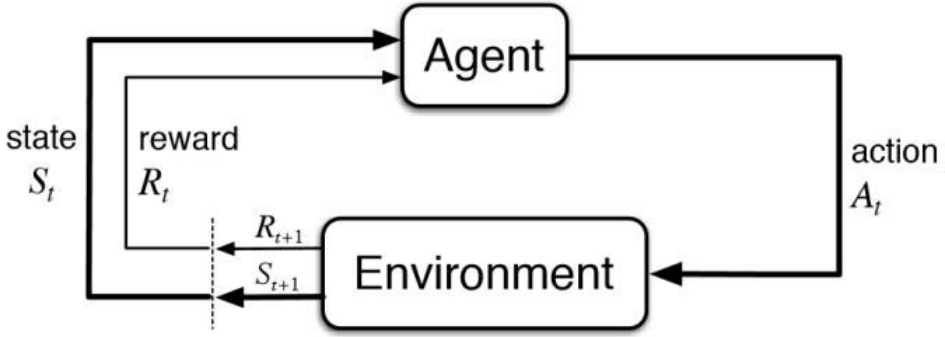
Apprentissage Automatique







Reinforcement Learning



L’agent doit apprendre en temps réel et sur place les séquences d’actions qui maximisent l’espérance des récompenses futures.

Applications possibles : contrôle d’un aéronef, drone, satellite, robot, ou agent virtuel.

Markovian Decision Process (MDP)

$$M = \begin{cases} S = \{s_1, s_2, ..., s_n\} & \text{is the finite set of states} \\ A = \{a_1, a_2, ..., a_m\} & \text{is the finite set of actions} \\ P = \Pr(s'|s, a) & \text{is the transition function} \\ R = \Pr(r|s, a, s') & \text{is the reward function} \end{cases}$$

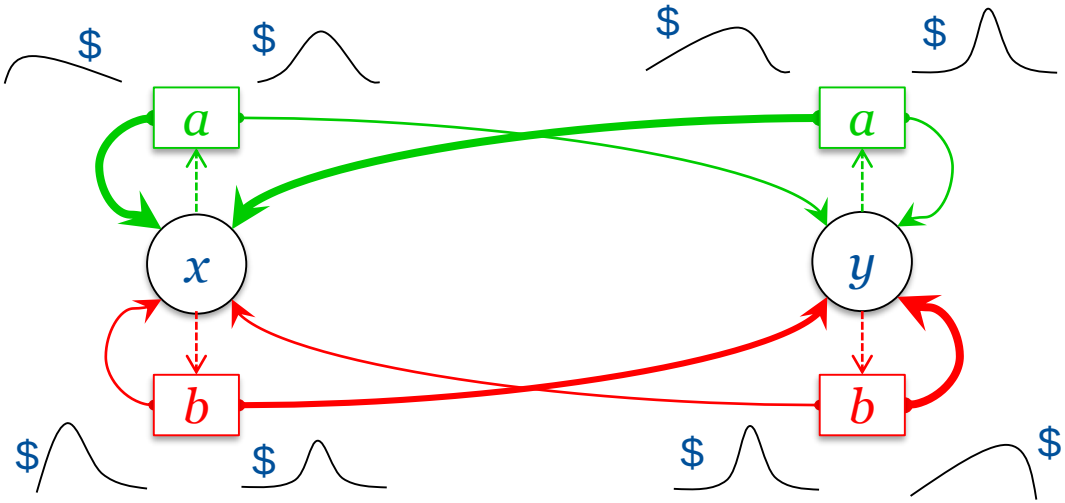
Solution : **a deterministic policy of actions**

$$\pi : S \rightarrow A$$

which maximizes the expected future rewards

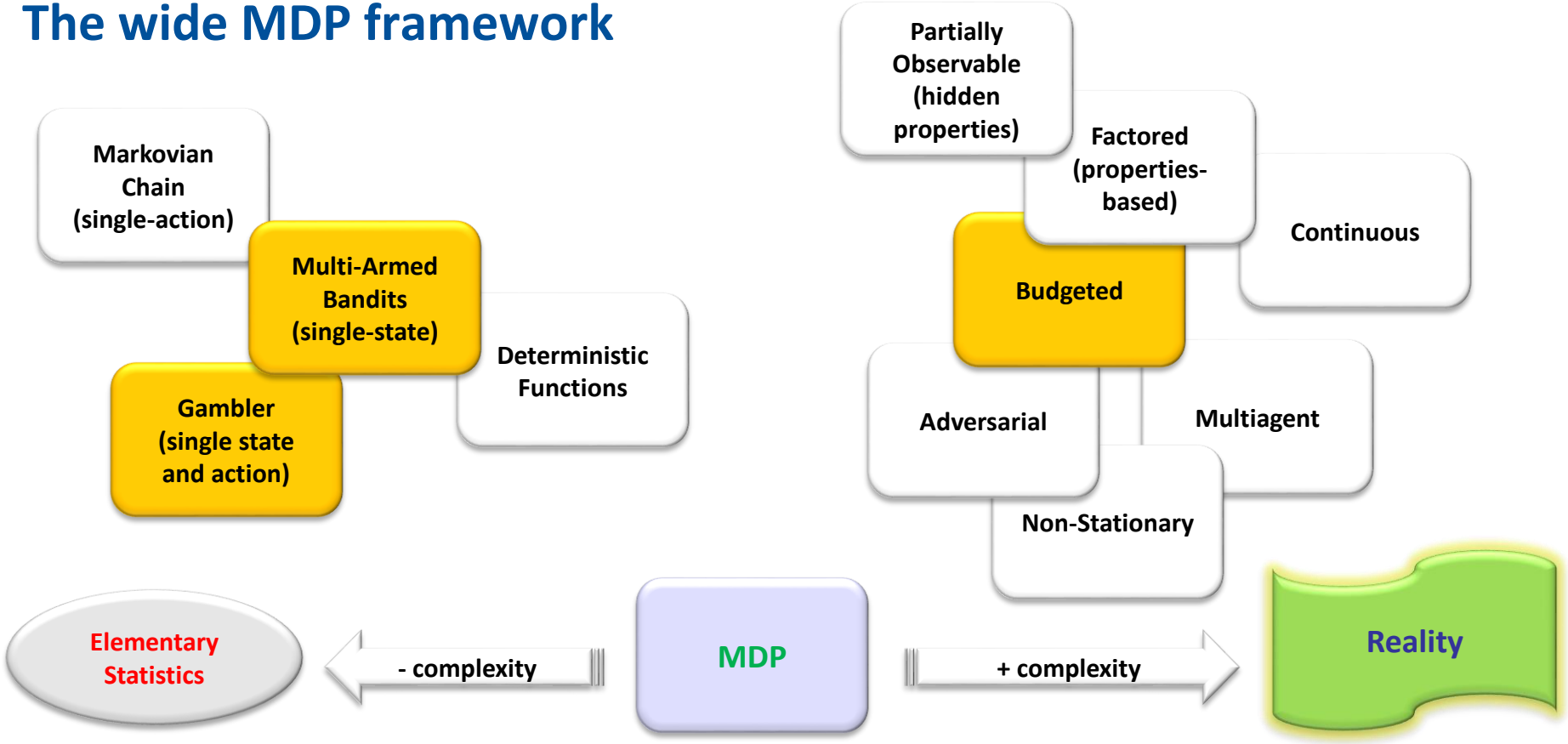
$$V_\gamma^\pi(s) = \lim_{k \to \infty} \sum_{t=1}^{k} \gamma^{t-1} R_t^\pi(s) \qquad \{\gamma \in \mathbb{R} \mid 0 \leq \gamma < 1\}$$

Processus de Décision Markovien



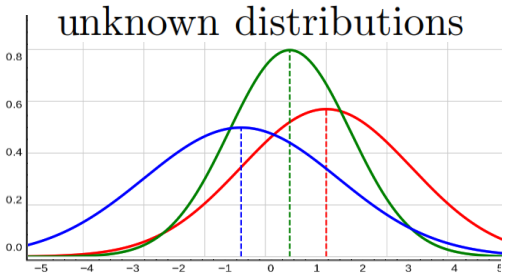
- **Reinforcement Learning : MDP is underlying – trial and error**
- **Dynamic Programming : MDP is given – exact solution**

The wide MDP framework



Standard Stochastic MAB

$$\mathcal{M} = \begin{cases} I = \{1, ..., k\} & \text{the set of possible actions} \\ F = \{f_1, ..., f_k\} & \text{the set of reward distribution functions} \end{cases}$$



exploration
(sampling arms)

exploitation
(pulling the estimated best arm)

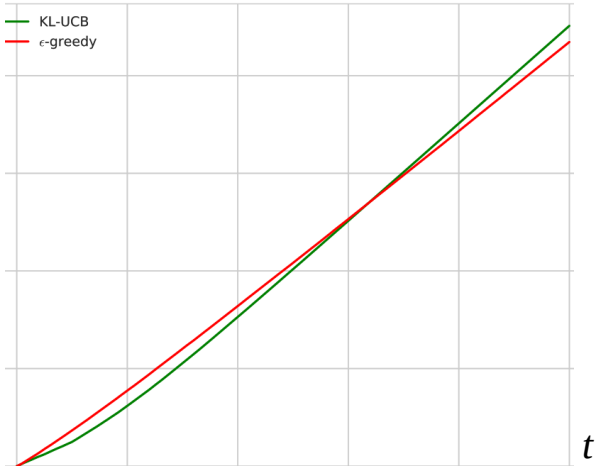
Standard MAB Optimization

goal : minimize the *regret*

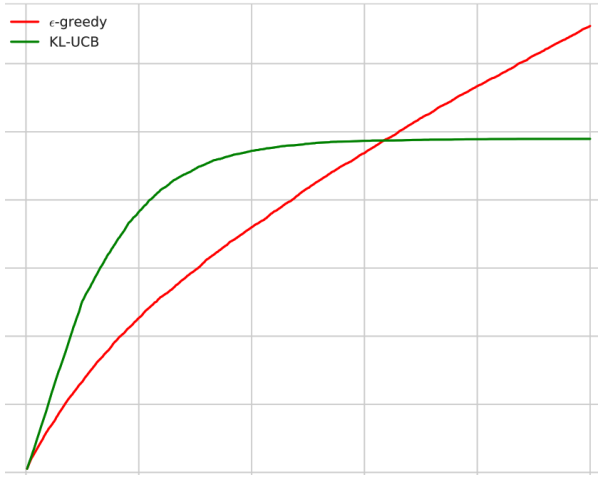
maximize the expected future rewards

$$\hat{\lambda}_h = \sum_{t=1}^{h} [\mu^* - \mu_{a_t}]$$

Cumulated Rewards



Cumulated Regret



Survival MAB

⇨ *exploration-exploitation-safety dilemma*

budget :

$$b_h = b_0 + \sum_{t=1}^{h} r_t \quad \rightarrow \quad b_{t+1} = b_t + r_t$$

the initial budget b_0 evolves in time following the received rewards

risk of ruin :

$$r_{min} < 0 < r_{max}$$

reward can just as well be positive as negative

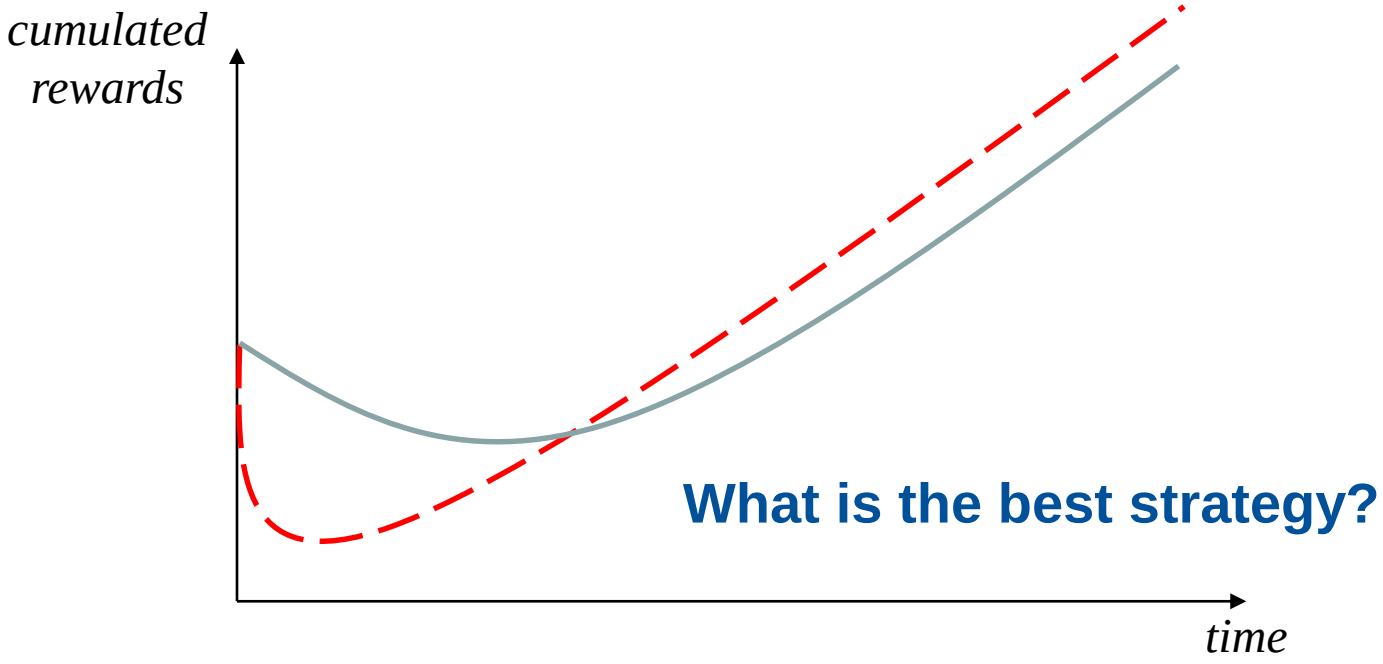
When the budget is over, the game is over.

multiobjective optimization:

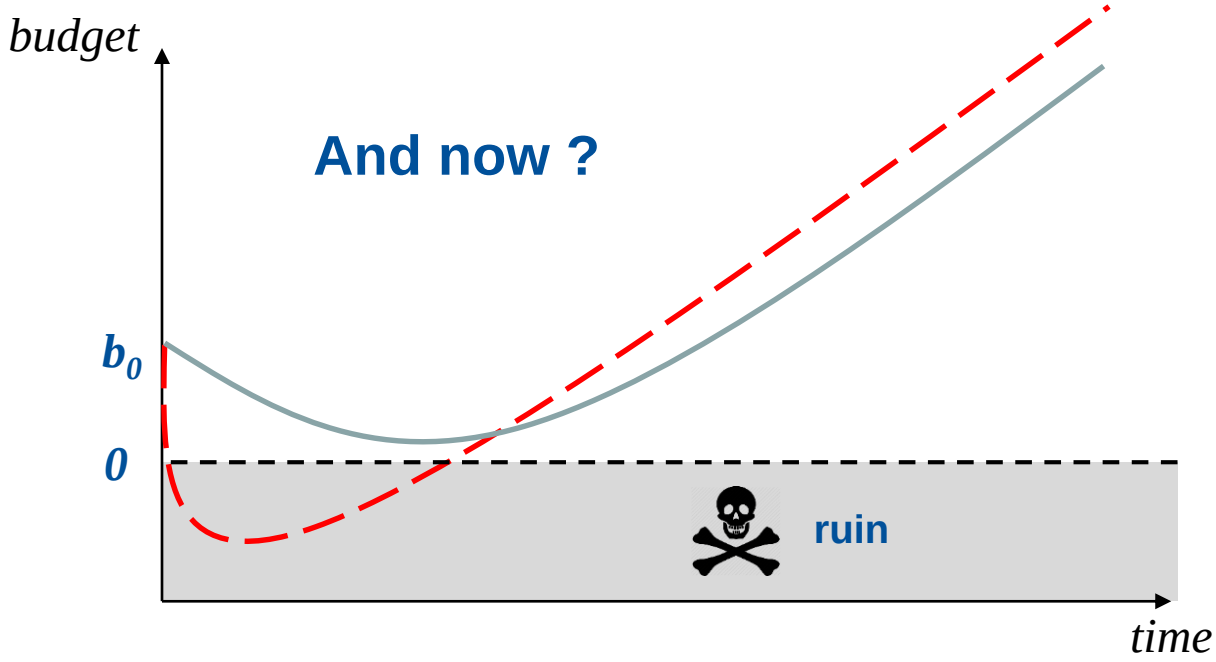
minimize the expected regret

while minimizing the probability of being ruined

Intuition : Cumulated Rewards



Intuition : Budget + Risk of Ruin



Perotto et al. (2021). AAMAS :

MODIFIED PROBLEM

A *multiarmed gambler bandit* (MAGB) is a random process that exposes $k \in \mathbb{N}^+$ arms to an agent having an initial budget $b_0 \in \mathbb{N}^+$, which evolves in time with the received rewards:

$$B_h = b_0 + \sum_{t=1}^{h} R_t$$

Let $\mathcal{P} = \{p_1, \cdots, p_k\}$ be the set of parameters that regulate the underlying Bernoulli distributions from which the rewards $R_t \in \{+1, -1\}$ are drawn.

At each round $t \in \mathbb{N}^+$, the agent executes an action i, which either increases its budget B_t by 1 with stationary probability $p_i \in [0, 1]$, or decreases it by 1 with probability $1 - p_i$.

The game stops when $B_t = 0$ happens for the first time (the gambler is ruined), but it can be occasionally played forever if the initial conditions allow the budget to increase infinitely.

The probability of surviving, never being ruined, having a current budget B_t, and repeatedly pulling arm i, is:

$$\lim_{h\to\infty} \omega_{h,i} = \begin{cases} 1 - \left(\frac{1-p_i}{p_i}\right)^{B_t} & \text{if } p_i > 0.5\,, \\ 0 & \text{if } p_i \leq 0.5\,. \end{cases}$$

PROPOSED METRIC

In contrast to the standard MAB, solving a MAGB involves a multi-objective optimization: in addition to minimizing the expected regret generated by the rounds when the best arm is not played (classic regret), the agent must also minimize the expected regret generated by the probability of being ruined.

To analyze that, we define the notion of *expected normalized relative regret* $\ell \in [0, 1]$:

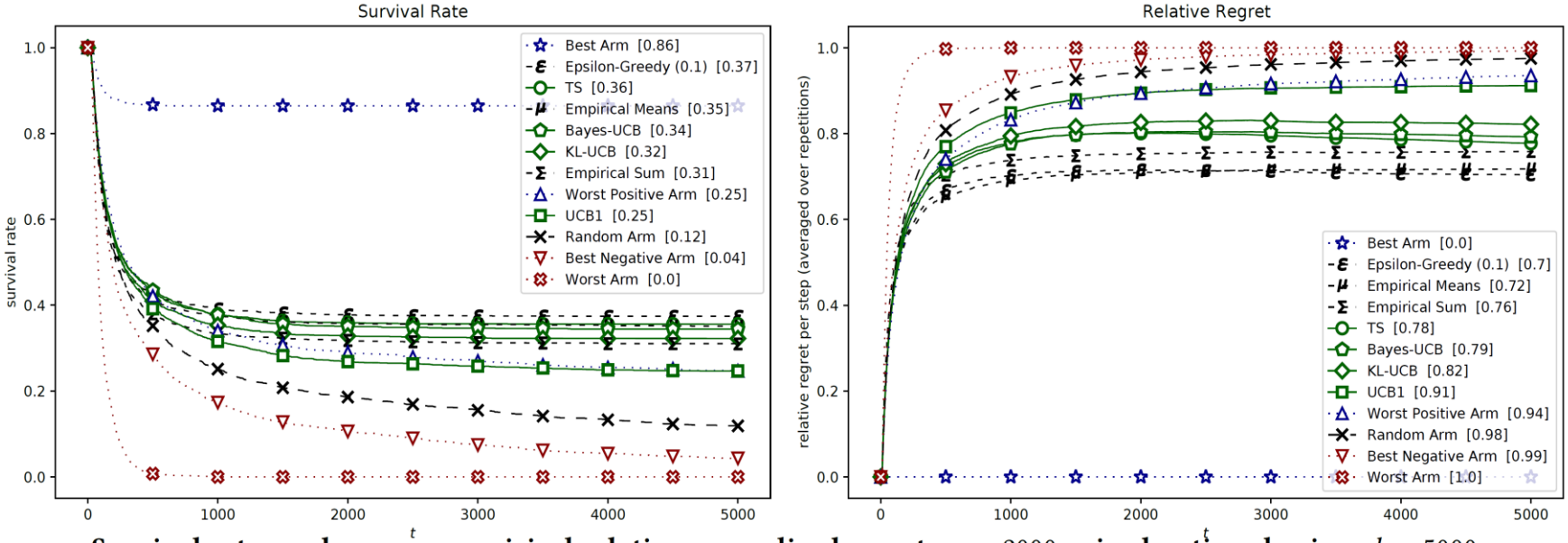
$$\ell_{h,\pi} = \frac{\omega_{h,\pi}}{\omega_h^*} \cdot \underbrace{\sum_{i=1}^{k} \left[\frac{p^* - p_i}{p^*} \cdot \frac{\mathbb{E}[N_{i,h}]}{h}\right]}_{\text{normalized classic regret}} + \underbrace{\left(\frac{\omega_h^* - \omega_{h,\pi}}{\omega_h^*}\right)}_{\text{regret due to ruin}},$$

where h is the considered (potentially infinite) time-horizon, p^* and p_i are, respectively, the underlying parameters of the optimal arm and of arm i, $\mathbb{E}[N_{i,h}]$ is the number of rounds arm i is expected to be pulled, and $\omega_{h,\pi}$ and ω_h^* are the probability of surviving, respectively, following a given strategy π, or always playing the best arm.

In finite-horizon experimental scenarios, after several independent repetitions, the expected normalized relative regret can be approximated empirically by averaging the normalized difference between the obtained final budget and the potentially best budget:

$$\hat{\ell}_{h,\pi} = 1 - B_{h,\pi}/B_h^*\,.$$

Perotto et al. (2021). AAMAS :



Survival rates and average empirical relative normalized regrets, $n = 2000$ episodes, time-horizon $h = 5000$.

Heuristic : Safety Threshold

- Chose your prefered classic MAB algorithm
- Tune the parameter ω *(safety threshold)*

1: **if** $b_t < \omega$ **and** $\max_i[\hat{\mu}_{i,t}] > 0$ **then**
2: $\quad a_t \leftarrow \arg\max_i[\hat{\mu}_{i,t}]$ *{gourmand}*
3: **else**
4: $\quad a_t \leftarrow$ **call** $\mathcal{A}$ *{méthode base}*
5: **end if**

Budget is too low, try to save yourself!
- if you know a positve arm, pull it.

Positive Gambler UCB

UCB:

estimated mean + confidence margin

log increased with time

$$\hat{\mu}_{i,t} + \sqrt{\frac{2 \ln t}{n_{i,t}}}$$

PG-UCB:

$$P^+_{i,t} + \sqrt{\frac{2 \ln b_t}{n_{i,t}}}$$

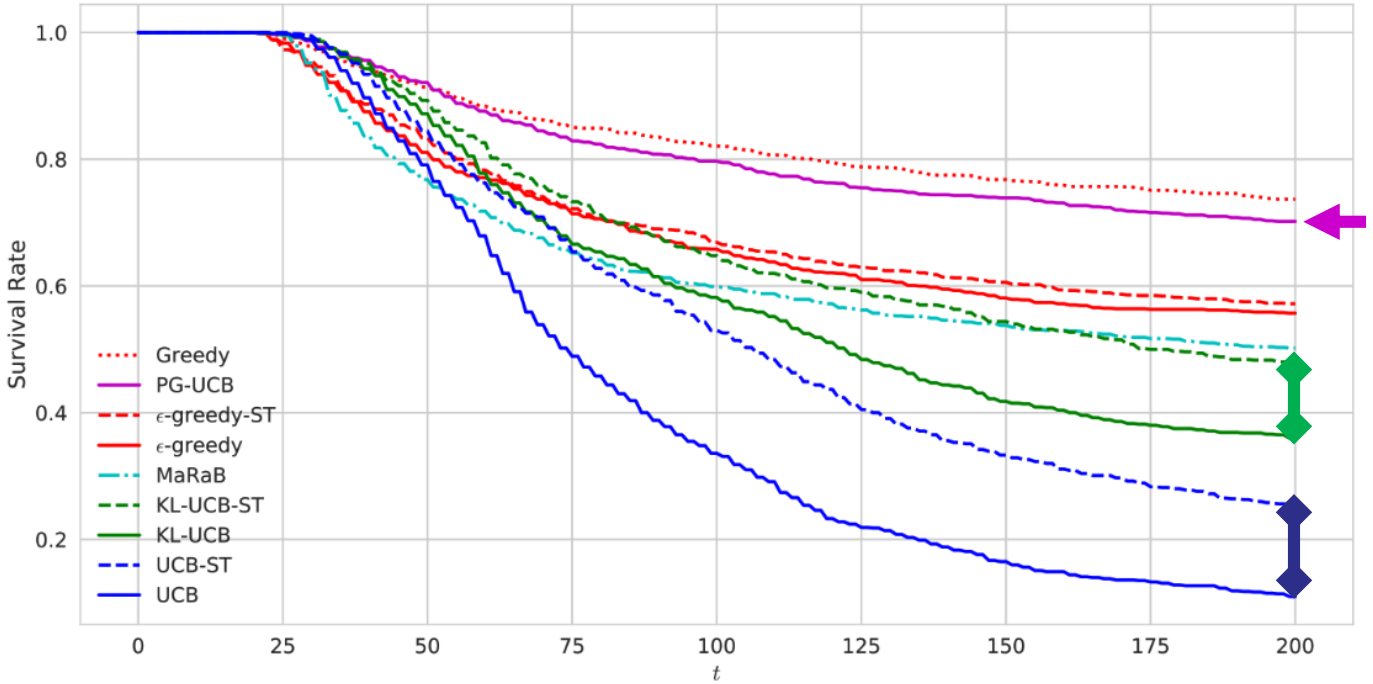
log increased with the budget

probability of having a positive mean

$$P^+ = P(\mu \geq 0 \mid \hat{\mu}, n) = \begin{cases} \theta/2 & \text{if } \hat{\mu} \leq 0 \\ 1 - \theta/2 & \text{if } \hat{\mu} > 0 \end{cases}$$

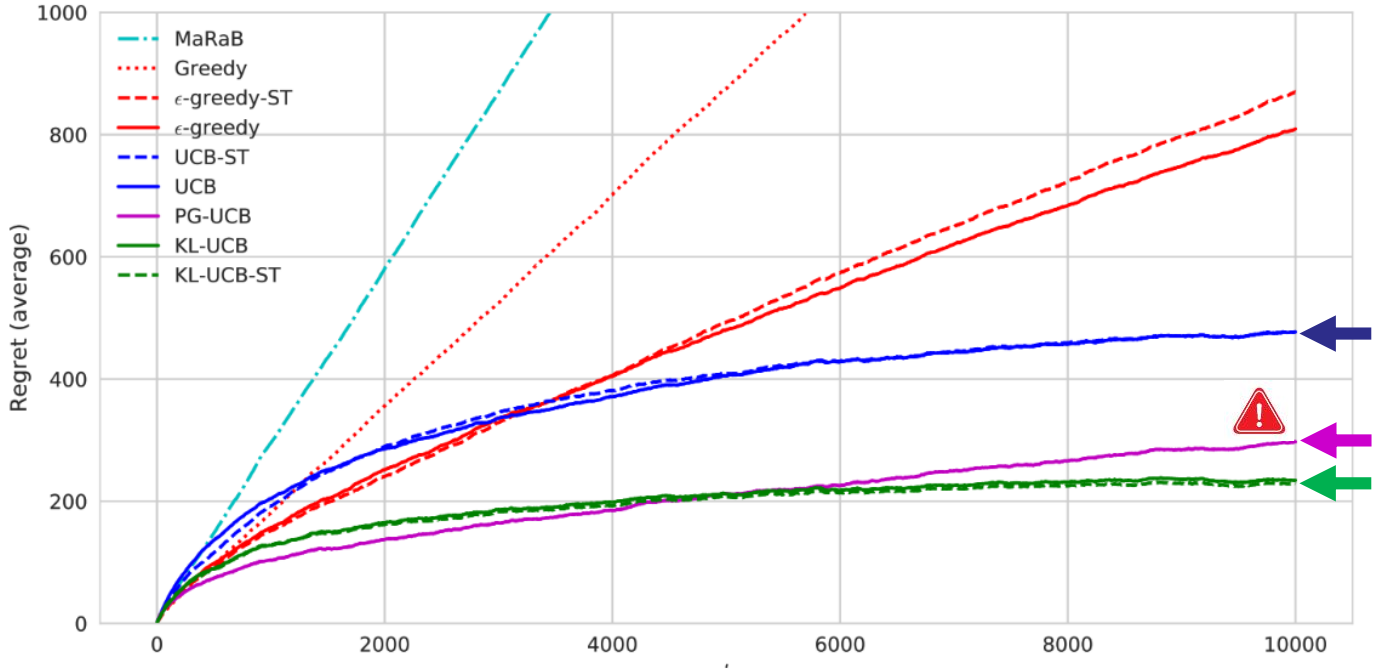
$$\theta = \exp\left(\frac{2n\hat{\mu}^2}{\Delta_r^2}\right) \qquad \Delta_r = r_{max} - r_{min}$$

Results : Survival Rate



10 arms, r = {+1, -1}, p = { .0 , …. , .6 }, $b_0 = 20$

Results : Regret



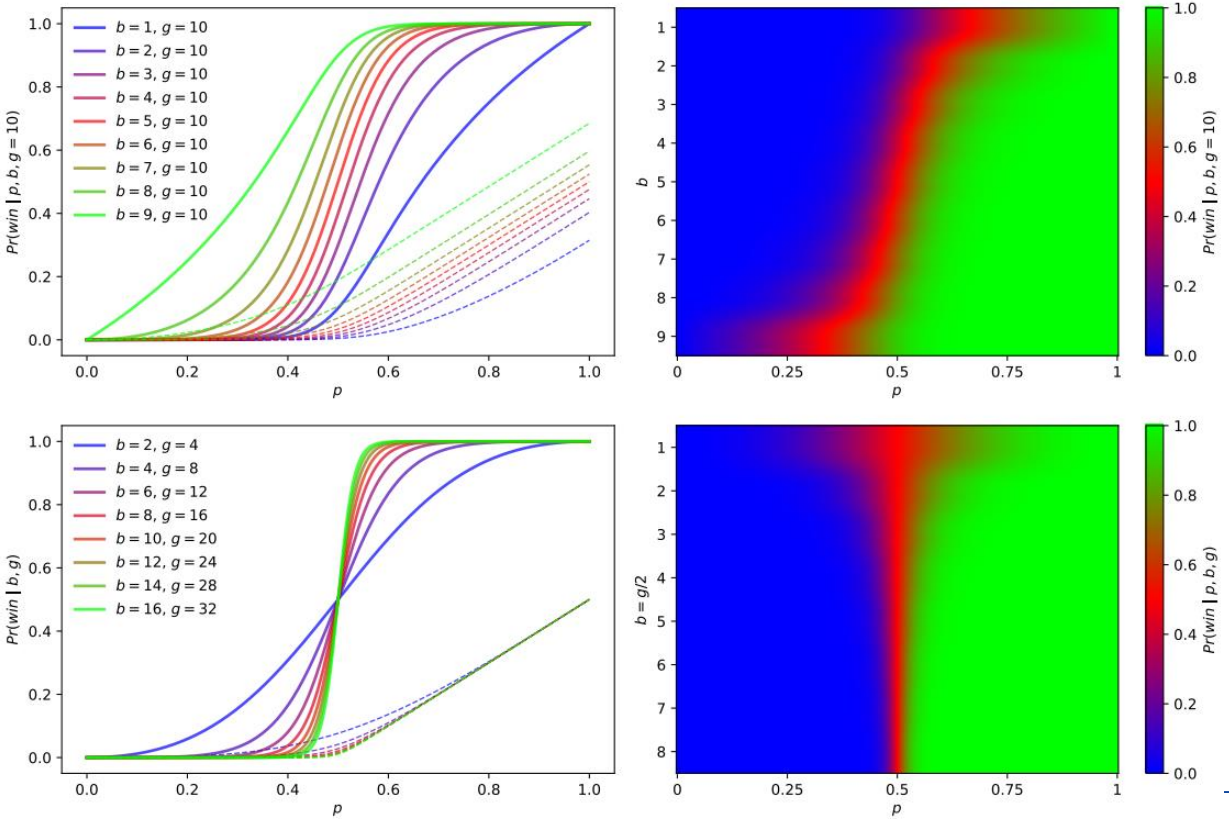
10 arms, r = {+1, -1}, p = { .0 , …. , .6 }, $b_0 = 20$

Classic Gambler's Ruin *(letter from P. de Fermat to B. Pascal)*

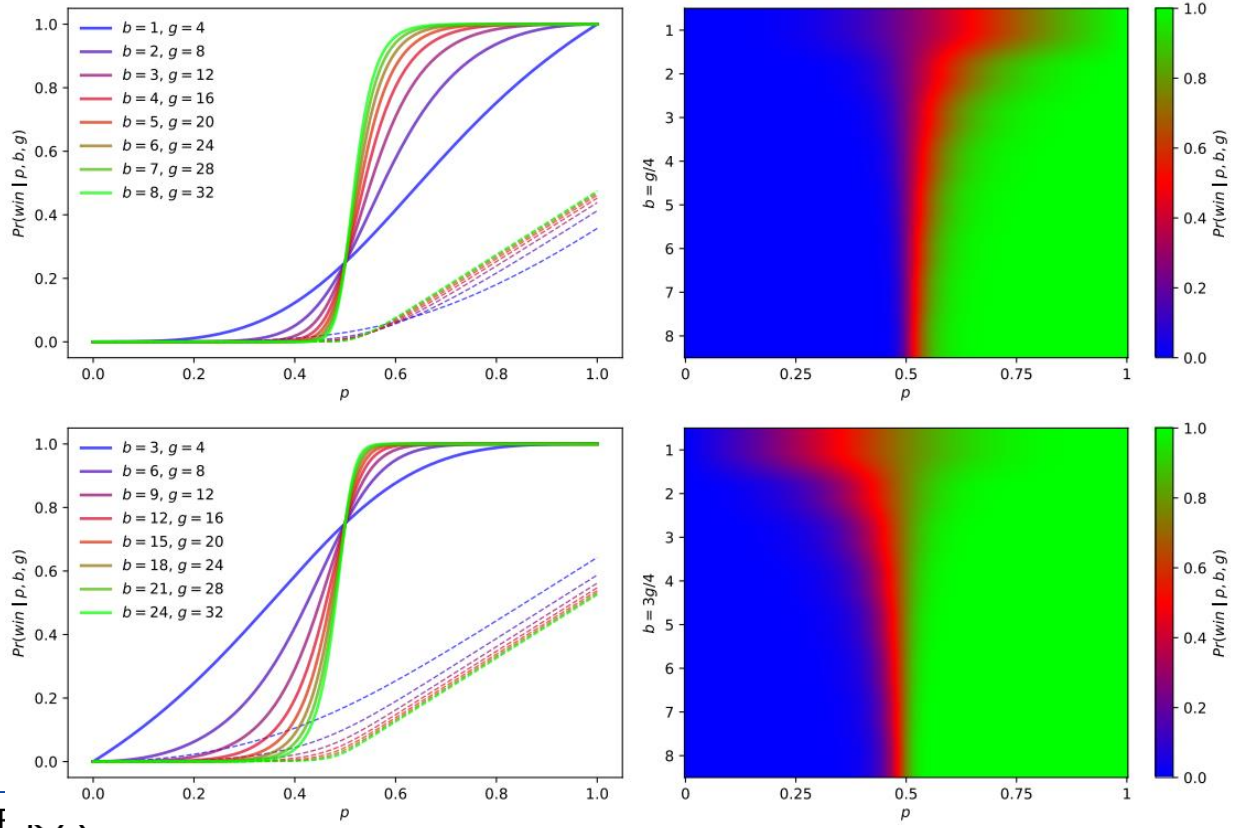
- The gambler starts with budget b_0
- At each round she/he can
  - get 1 with probability p
  - or loss 1 with probability $q = 1–p$
- The gambler wins if she/he reaches g
- And is ruined if the budget is over

$$\mathbb{P}^{win}_{bold}(b_t, g, p) = \mathbb{P}(s_\tau = win \mid b_t, g, p, \pi_{bold}) = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^{b_t}}{1-\left(\frac{q}{p}\right)^{g}} & \text{if } p \neq \frac{1}{2}, \\ \frac{b_t}{g} & \text{if } p = q = \frac{1}{2}. \end{cases}$$

Gambler's Ruin: Probability of Winning



Gambler's Ruin: Probability of Winning



Perotto et al. (2021). IJAR

- At each round, the agent can decide to stop playing keeping earnings.

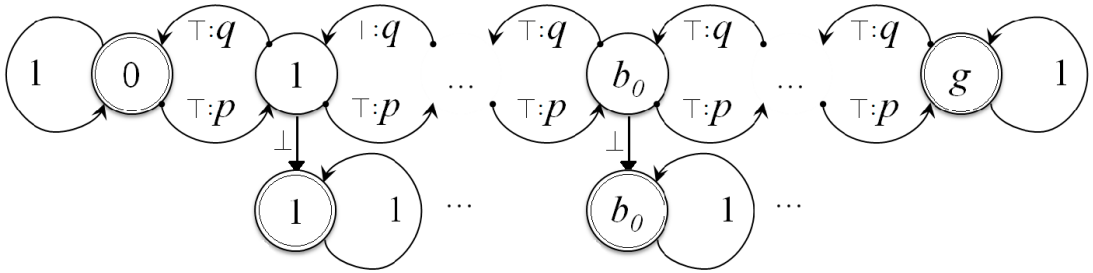


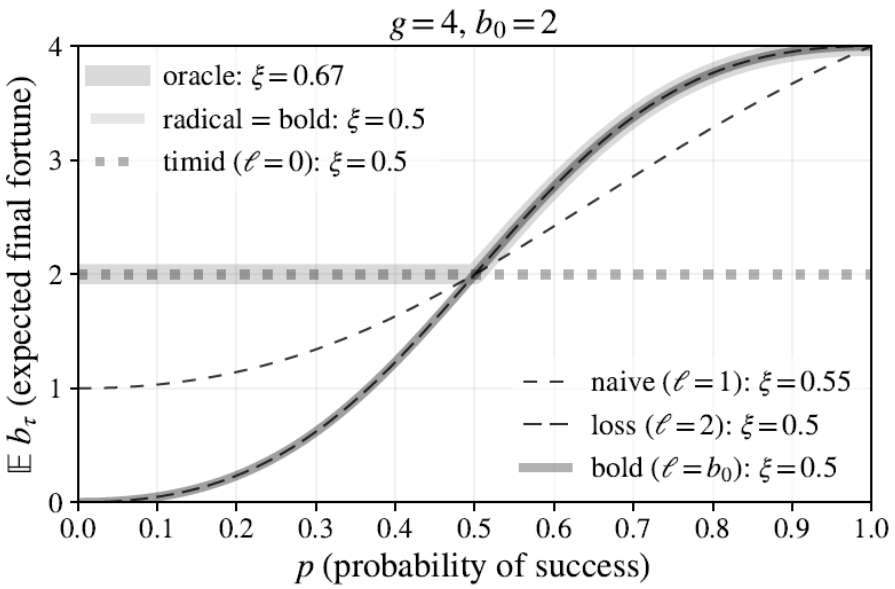
Figure 1: A *decisional gambling process* can be modeled as a *Markovian decision process*. Each node in the top of the graph corresponds to a possible budget while the game is running. The nodes 0 and g are absorbing (self-loop with probability 1), corresponding, respectively, to the *losing* and *winning* situations. The nodes in the bottom of the graph are also absorbing, corresponding to stopping the game with an intermediate budget, following the agent's decision to quit the game. The initial state is b_0.

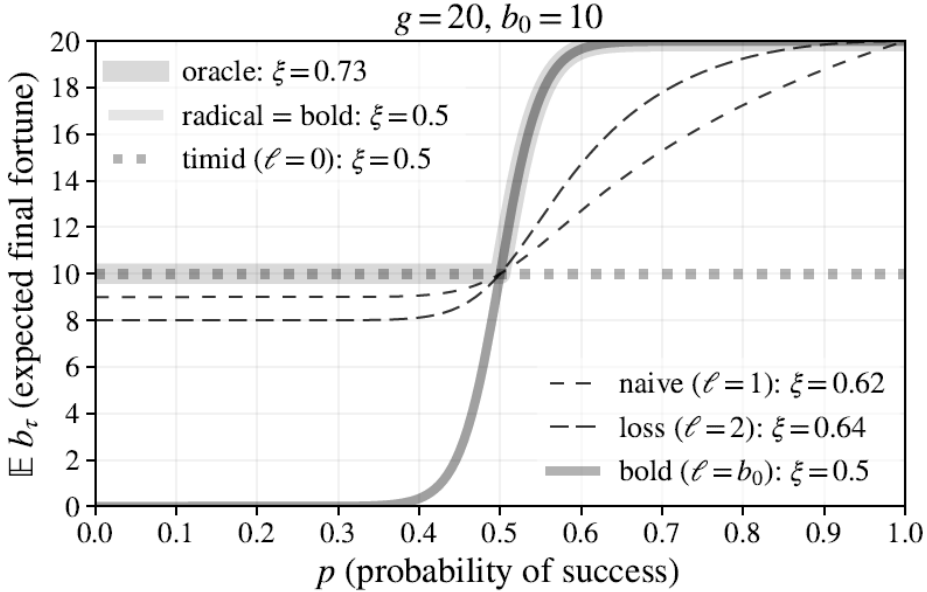
Possible Strategies:

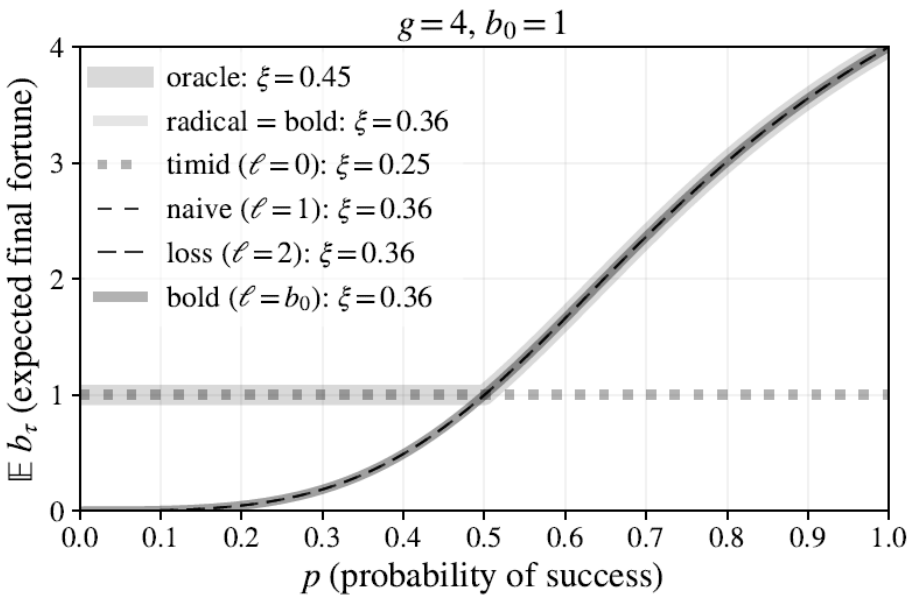
- Bold : everything or nothing, play always
- Timid : quit the game immediatelly
- Radical : bold if $b < g/2$, timid otherwise
- Oracle : bold if $p > 0.5$, timid otherwise
- Naive : play while estimated $p > 0.5$
- Loss-n : play while $b > b_0 - n$

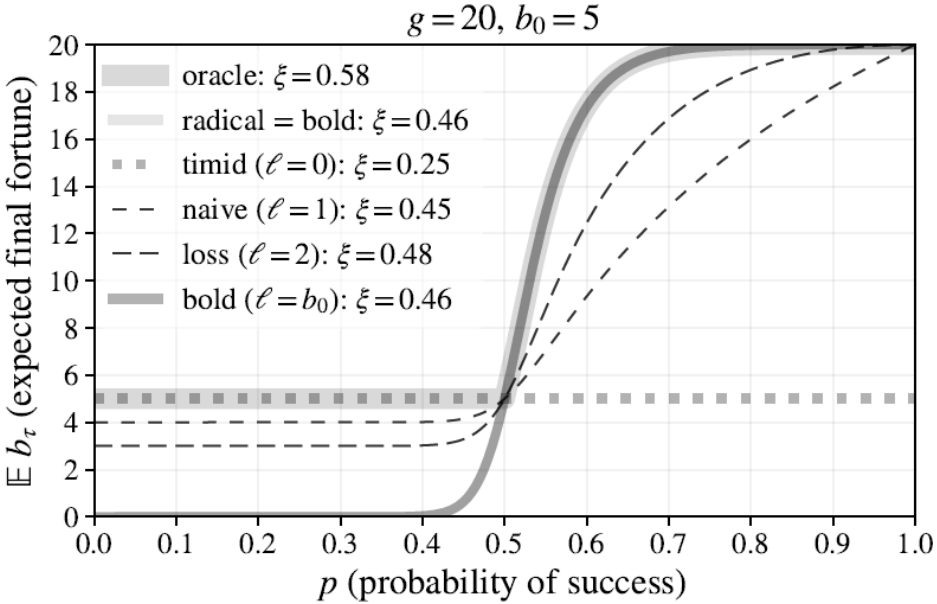
Possible Strategies:

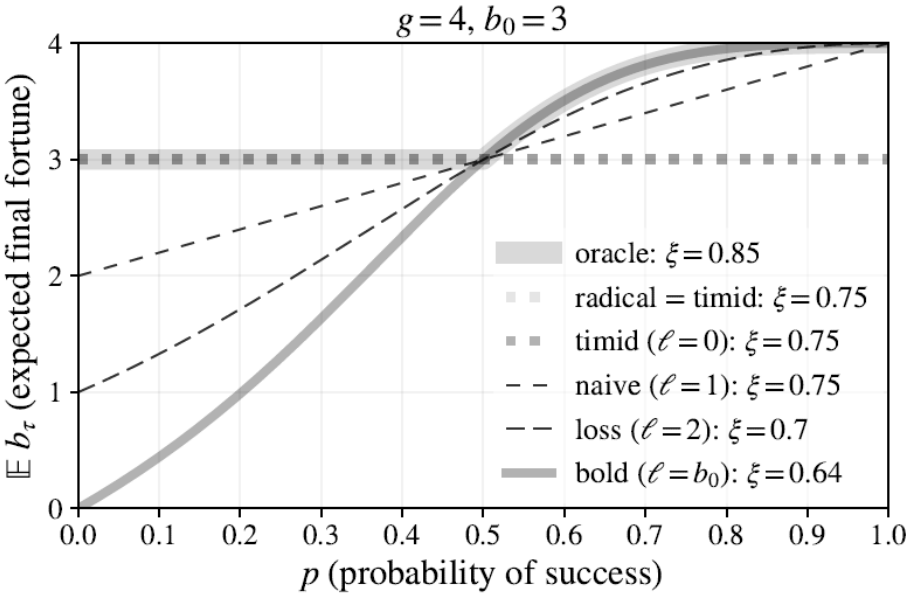
- Bold = Loss-b_0
- Timid = Loss-0
- Naive = Loss-1

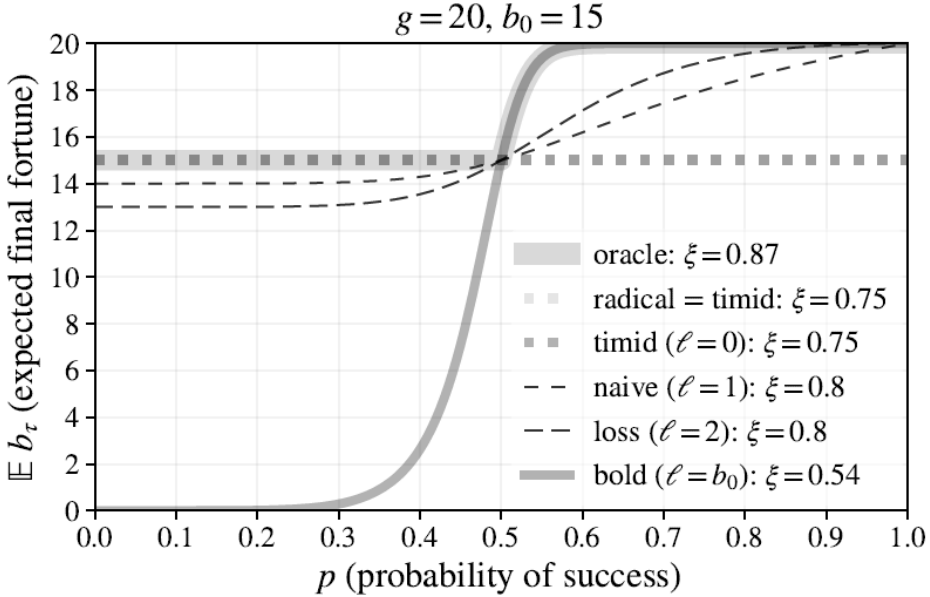

$g = 4, b_0 = 2$
oracle: $\xi = 0.67$
radical = bold: $\xi = 0.5$
timid ($\ell = 0$): $\xi = 0.5$
naive ($\ell = 1$): $\xi = 0.55$
loss ($\ell = 2$): $\xi = 0.5$
bold ($\ell = b_0$): $\xi = 0.5$
$\mathbb{E}\, b_\tau$ (expected final fortune)
p (probability of success)


$g = 20, b_0 = 10$
oracle: $\xi = 0.73$
radical = bold: $\xi = 0.5$
timid ($\ell = 0$): $\xi = 0.5$
naive ($\ell = 1$): $\xi = 0.62$
loss ($\ell = 2$): $\xi = 0.64$
bold ($\ell = b_0$): $\xi = 0.5$
$\mathbb{E}\, b_\tau$ (expected final fortune)
p (probability of success)

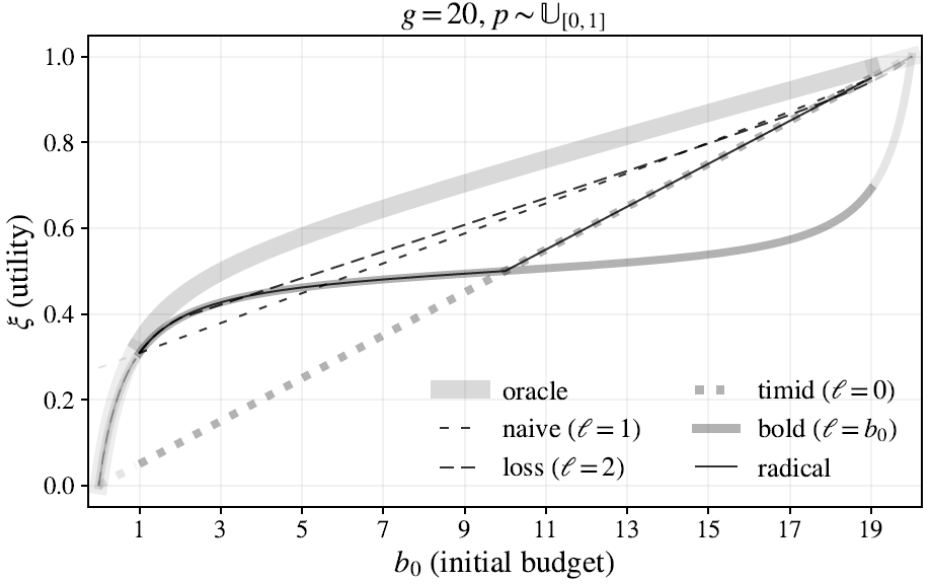

$g = 4,\ b_0 = 1$
oracle: $\xi = 0.45$
radical = bold: $\xi = 0.36$
timid ($\ell = 0$): $\xi = 0.25$
naive ($\ell = 1$): $\xi = 0.36$
loss ($\ell = 2$): $\xi = 0.36$
bold ($\ell = b_0$): $\xi = 0.36$
$\mathbb{E}\ b_\tau$ (expected final fortune)
p (probability of success)

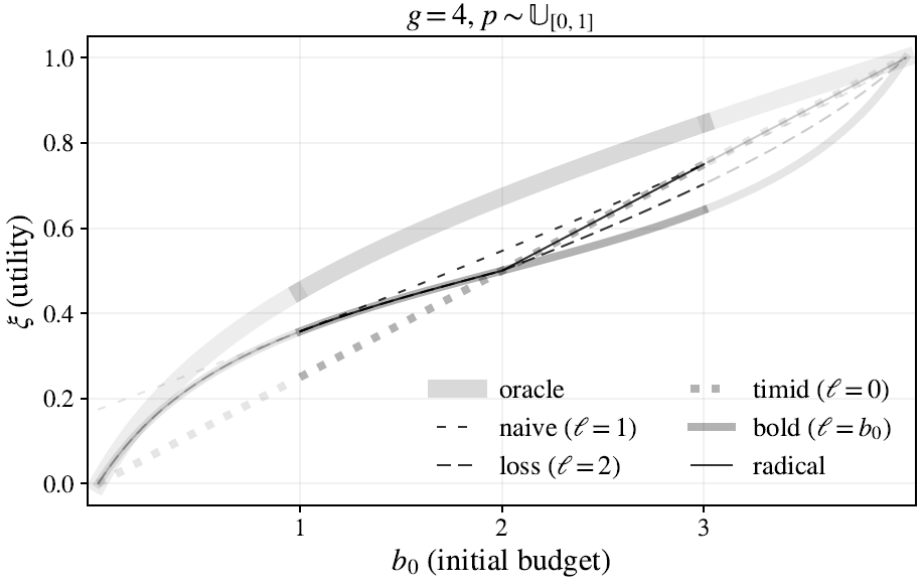

$g = 20,\ b_0 = 5$
oracle: $\xi = 0.58$
radical = bold: $\xi = 0.46$
timid ($\ell = 0$): $\xi = 0.25$
naive ($\ell = 1$): $\xi = 0.45$
loss ($\ell = 2$): $\xi = 0.48$
bold ($\ell = b_0$): $\xi = 0.46$
$\mathbb{E}\ b_\tau$ (expected final fortune)
p (probability of success)


$g = 4, b_0 = 3$
$\mathbb{E}\, b_\tau$ (expected final fortune)
p (probability of success)
oracle: $\xi = 0.85$
radical = timid: $\xi = 0.75$
timid ($\ell = 0$): $\xi = 0.75$
naive ($\ell = 1$): $\xi = 0.75$
loss ($\ell = 2$): $\xi = 0.7$
bold ($\ell = b_0$): $\xi = 0.64$


$g = 20, b_0 = 15$
$\mathbb{E}\, b_\tau$ (expected final fortune)
p (probability of success)
oracle: $\xi = 0.87$
radical = timid: $\xi = 0.75$
timid ($\ell = 0$): $\xi = 0.75$
naive ($\ell = 1$): $\xi = 0.8$
loss ($\ell = 2$): $\xi = 0.8$
bold ($\ell = b_0$): $\xi = 0.54$

Theoretical Comparison





$$\ell^* \mid g, b_0 \;=\; \max\left(1\,,\; \min\left(b_0\,,\; \left|\frac{-1+\sqrt{2g-2b_0-1}}{2}\right| + 1\right)\right)$$

Projet RG

- Apprentissage par Renforcement sous Risque de Ruine (A3R)
  - consolider résultats théoriques sur MAB
  - extension du travail sur les MDPs
  - possibles contributions aérospatiales

RÉPUBLIQUE
FRANÇAISE

Liberté
Égalité
Fraternité

ONERA

THE FRENCH AEROSPACE LAB

www.onera.fr