Mathematical Institute

Multi-armed bandits with uncertainty

Samuel N. Cohen
Mathematical Institute
University of Oxford

Based on joint work with Tanut Treetanthiploet

Research Supported by:
The Alan Turing Institute

Oxford Mathematics

In many settings, we have to make decisions with limited information

- This is common in finance, where estimation and model error is significant, but we have to act nevertheless
- In some cases, our actions will determine how our learning changes over time.
- If we care about making good decisions, both now and in the future, how should we choose what to do?

A classical example of this is a multi-armed bandit.

- One has M machines which you can play, but you do not know the distribution of their costs.
- When you play, you face a (random) cost, but will also learn about this distribution, which allows you to make better future decisions.
- You need to decide how to trade-off between *exploring*, that is, playing machines for the purposes of learning about costs, and *exploiting*, that is, playing to achieve a small cost.

These are used as toy models of many different problems: drug trials, trading strategies, optimal scheduling, experimental design...

The analysis of these systems has a long history.

- The problem was first formulated during WWII, and according to Peter Whittle "efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate instrument of intellectual sabotage."
- A great advance was made by Gittins (Gittins & Jones, 1974),
  - Suppose the payoffs depend on an observed Markov process (for example, the Bayesian estimates) and are independent
  - He constructed a 'dynamic allocation index' (the *Gittins' index*) which is a predictable process such that the minimal-expected-cost strategy is to play the machine with the smallest index.

Let's try and make Gittins' result precise

- We have costs $h^{(\rho_t)}(\omega, t)$
  - These can depend on observed 'state' variables.
  - ρ_t indicates which machine to play at time t.
- (Weber) If $\gamma_t^{(m)}$ is the $\mathcal{F}_t$-measurable value such that

$$0 = \sup_{\tau} E\Big[\sum_{t<s\leq\tau} \beta^s \big(h^{(m)}(\omega, s) - \gamma_t^{(m)}\big) \Big| \mathcal{F}_t \Big] \qquad \text{for } \tau > 0 \text{ stopping times.}$$

- The strategy $\rho_t = arg\ min_m\{\gamma_t^{(m)}\}$ minimizes

$$E\Big[\sum_{t<s\leq\tau} \beta^s h^{(\rho_s)}(\omega, s) \Big| \mathcal{F}_t \Big].$$

Other approaches exist

- Mandelbaum proposes an allocation strategy approach
- Whittle uses a dynamic programming argument based on a 'retirement option'
- Numerous extensions, variations, ...
- Extended to continuous time by El Karoui and Karatzas, Bank and Föllmer, Bank and Küchler,...

Effectively all treat only the expected cost criterion or a (linear) regret criterion.

- Kelly (see also Bather, Chang and Lai, Yao...) noticed that if $\beta \to 1$, in a simple case, the rule degenerates towards the following rule: "Choose the bandit where you've seen the least number of losses."
- Equivalently, take a confidence bound on the expected return of each bandit, and play the bandit with the lowest confidence bound
- In the CS community, Agrawal (and others) proposed the 'UCB' algorithm, by tweaking the width of the confidence intervals, and proved asymptotic performance bounds.

This leads to a peculiar conclusion:

You should prefer to play uncertain machines.

- Model uncertainty, where you don't assume you know the distribution of outcomes, has been of much interest in the mathematical finance community.
  - Classic works include Knight (1921), Keynes (1921) and Wald (1950)
  - Much recent work uses sets of probability measures or nonlinear expectations/risk measures to describe uncertainty
- Paradoxes such as the Ellsberg paradox suggest that we are generally biased *against* uncertainty
  - This is easily checked by introspection
- Connections with statistical estimation are possible

Example

- Suppose we are going to bet on the toss of a (possibly unfair) coin.
- Heads you get £1, tails you get nothing.
- You must choose between two possible coins.

Example

- Suppose we are going to bet on the toss of a (possibly unfair) coin.
- Heads you get £1, tails you get nothing.
- You must choose between two possible coins.
- We will throw each coin i exactly N_i times before you choose.
- Given your observations, which coin do you choose?

Example (Frequentist solution (?))

- The first we throw $N_1 = 3$ times, and observe 2 heads
- The second we throw $N_2 = 3000$ times, and observe 2000 heads.
- Which coin do you prefer?

Example (Frequentist solution (?))

- The first we throw $N_1 = 3$ times, and observe 2 heads
- The second we throw $N_2 = 3000$ times, and observe 2000 heads.
- Which coin do you prefer?
- Everyone prefers the second. Most people still prefer the second if we only observe 1999 heads.
- This is inconsistent with the basic rule: maximize the estimated expectation $\hat{E}[u(X^i)] = \hat{p}_i u(1)$.

Example (Bayesian solution (?))

- For any prior distribution on p, we can compute

$$E[u(X^i)|\text{obs}] = E\big[E[u(X^i)|p_i, \text{obs}]\big|\text{obs}\big] = E[p_i|\text{obs}]u(1)$$

- The variance of p doesn't appear in the Bayesian value
- The utility/loss function u adds nothing
- The prior has effect $O(1/N)$, while estimation error is $O(1/\sqrt{N})$.

Given a space $(\Omega, \mathcal{F}, \mathbb{P})$, a coherent (nonlinear) expectation is a map $\mathcal{E} : L^\infty(\mathcal{F}) \to \mathbb{R}$ such that

- Monotonicity: If $\xi \geq \xi'$ $\mathbb{P}$-a.s. then $\mathcal{E}(\xi) \geq \mathcal{E}(\xi')$
- Constant triviality and equivariance: For any $k \in \mathbb{R}$, $\xi \in L^\infty$, $\mathcal{E}(k) = k$ and $\mathcal{E}(\xi + k) = \mathcal{E}(\xi) + k$
- Positive homogeneity: for $\lambda > 0$, $\mathcal{E}(\lambda\xi) = \lambda\mathcal{E}(\xi)$
- Subadditivity: $\mathcal{E}(\xi + \xi') \leq \mathcal{E}(\xi) + \mathcal{E}(\xi')$
- Lebesgue property: If $\xi_n \to \xi$ then $\mathcal{E}(\xi_n) \to \mathcal{E}(\xi)$.

The measure $\mathbb{P}$ acts as a 'reference' measure, and we assume strict monotonicity $\mathbb{P}$-a.s.

We can describe a coherent expectation through its dual (see Artzner et al.):

- Under our assumptions $\mathcal{E}$ has the representation

$$\mathcal{E}(\xi) = \sup_{Q \in \mathcal{Q}} \left\{ E_Q[\xi] \right\}$$

An example of a family of measures $\mathcal{Q}$ (over a single step) is given by confidence intervals, or balls around an estimate in Wasserstein distance.

- Note that if ξ represents losses, our preference is for *less* uncertain options.
- What happens when we put this theory into practice for a multi-armed bandit?

A key problem when working with nonlinear expectations is time consistency.

- Suppose our information is modelled by a filtration $\{\mathcal{F}_t\}$.
- If expectations can be obtained through Backward induction/BSDEs, then they have recursivity: we can define $\mathcal{E}(\cdot|\mathcal{F}_t)$ such that

$$\mathcal{E}(\cdot|\mathcal{F}_t) = \mathcal{E}\big(\mathcal{E}(\cdot|\mathcal{F}_s)\big|\mathcal{F}_t\big) \qquad \text{for all } t < s.$$

- Coherent expectations are recursive for *all* filtrations iff they are linear (or the trivial worst-case expectation).
- Given recursivity, decisions made to maximize/minimize the expectation will satisfy a dynamic programming principle.

- In the multi-armed bandit problem, our controls determine the filtration.
- This means we cannot presume consistency in our expectation.
- The earlier optimality condition is then too strong to obtain results.
  - One could use backward induction over all possible observation sets to ensure consistency, but then you do not have a simple optimality criterion, and there is no hope of a 'nice' solution to the problem.
- Using an indifference valuation perspective can be helpful

- We have M bandits, each with a filtered space $(\Omega^{(m)}, \mathcal{F}^{(m)})$ and dominating measure $\mathbb{P}^{(m)}$. Assume we eventually play machine m exactly $T^{(m)}$ times.
- Define the *orthant space* by $\bar{\Omega} = \bigotimes_m \Omega^{(m)}$, similarly $\bar{\mathbb{P}}$, and

$$\bar{\mathcal{F}}(s) = \bigotimes_m \mathcal{F}^{(m)}(s^{(m)})$$

  where $s = (s^{(1)}, s^{(2)}, ..., s^{(m)})$.
- $\bar{\mathcal{F}}(s)$ is a filtration in the sense that if $s \leq s'$ componentwise then $\bar{\mathcal{F}}(s) \subseteq \bar{\mathcal{F}}(s')$
- A policy ρ induces a map $t \to s$ by counting the number of plays, so we can write $\mathcal{F}^{\rho}_t = \bar{\mathcal{F}}_{s[\rho,t]}$ for the 'observed' filtration.

- We are now ready to show that there exists an index process which gives an optimal strategy, in some sense.
- The proof depends heavily on the independence of the bandit processes
  - We assume our expectation is recursive on an individual Bandit
  - The bandits have a strong form of independence in their (uncertain) distributions
- Given this structure, we separate into two key steps: analysing a single Bandit and combining multiple Bandits together

- For each m, we have a family of measures $\mathcal{Q}^{(m)}$, with $\mathbb{Q} \ll \mathbb{P}^{(m)}$ for all $\mathbb{Q} \in \mathcal{Q}^{(m)}$.
- The set $\mathcal{Q}^{(m)}$ describes our (evolving) uncertain family of models for the mth bandit.
- We do not require stationarity or ergodicity assumptions, and Bayesian updating can be built into each model.
- Define $\bar{\mathcal{Q}} = \{\mathbb{Q} = \bigotimes_m \mathbb{Q}^{(m)} \text{ for } \mathbb{Q}^{(m)} \in \mathcal{Q}^{(m)}\}$.

- For each m, we assume $\mathcal{Q}^{(m)}$ satisfies a pasting property:
  If $\mathbb{Q}, \mathbb{Q}' \in \mathcal{Q}^{(m)}$ then for any s,

$$\hat{\mathbb{Q}}(A) := \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}'}[I_A|\mathcal{F}_s^{(m)}]\right]$$

  gives a measure in $\mathcal{Q}^{(m)}$
- Equivalently, the nonlinear expectation

$$\mathcal{E}^{(m)}(\xi|\mathcal{F}_s^{(m)}) = \sup_{\mathbb{Q}\in\mathcal{Q}^{(m)}} \mathbb{E}_{\mathbb{Q}}[\xi|\mathcal{F}_s^{(m)}]$$

  is recursive.
- In examples, we will force this by constructing $\mathcal{E}^{(m)}$ by backward induction

Theorem

Let $\mathfrak{E}_s(\xi) := \sup_{\mathbb{Q}\in\bar{\mathcal{Q}}} \mathbb{E}_{\mathbb{Q}}[\xi|\bar{\mathcal{F}}(s)]$. *Then* $\mathfrak{E}$ *satisfies*

- *Subconsistency:* $\mathfrak{E}_s(\xi) \leq \mathfrak{E}_s(\mathfrak{E}_{s'}(\xi))$ *for* $s \leq s'$
- *Independence: For nonnegative single-machine random variables* $X^{(m)} : \Omega^{(m)} \to \mathbb{R}_+$

$$\mathfrak{E}_s\Big(\prod_m X^{(m)}(\omega_m)\Big) = \prod_m \mathcal{E}^{(m)}\big(X^{(m)}(\omega_m)\big|\mathcal{F}^{(m)}_{s^{(m)}}\big).$$

We can also extend $\mathfrak{E}_s$ to 'stopping times' $\mathfrak{T}$, where $S \in \mathfrak{T}$ if $S = (S^{(1)}, S^{(2)}, ..., S^{(M)})$ and $S^{(m)}$ is an $(\mathcal{F}^{(m)}_s)_{s\geq 0}$-stopping time for each m.

Definition

We say a C^ρ is a compensator of a discounted cost process g^ρ if

$$\mathfrak{E}\Big(\sum_n (g_n^\rho - C_n^\rho)\Big) = 0$$

and C_n^ρ is $\mathcal{F}_{n-1}^\rho$-measurable.

We say ρ^* is a Gittins' optimum if there exists a compensator family $\{C^\rho\}$ such that, for all N,

$$\mathfrak{E}\Big(I_A \sum_{n>N} (g_n^{\rho^*} - C_n^{\rho^*})\Big) \le 0 \qquad \text{for all } A \in \mathcal{F}_N^{\rho^*}$$

and, for all $N \ge 1$, all ρ, $\sum_{n=1}^N C_n^{\rho^*} \le \sum_{n=1}^N C_n^\rho$ a.s.

- In a classical setting, writing $C_n^\rho = E[V_{n+1}^\rho | \mathcal{F}_n] - V_n^\rho$ gives such a family C^ρ, by the martingale optimality principle.
- Generally C^ρ is not unique but one can derive a backward induction algorithm to compute a possible C^ρ
  - The key idea is that you account for future (random) compensation when computing the minimal compensation required today
  - Our definition gives a Strotz–Pollack (intertemporal game) style time-consistency, at least in the optimal strategy
- With a recursive expectation, (classical) optimality implies Gittins' optimality, and conversely *for some predictable endowment*.

For each machine m, we have an $(\mathcal{F}_t^{(m)})_{t\geq 0}$-adapted sequence of random costs $h^{(m)}$. Assume (for simplicity) these generate $(\mathcal{F}_t^{(m)})_{t\geq 0}$.

Definition

For each m, define the *robust Gittins' index* $\gamma_t^{(m)}$ to be the smallest $\mathcal{F}_t^{(m)}$-measurable value γ such that

$$\operatorname*{ess\,inf}_{\tau\in\mathcal{T}^{(m)}} \mathcal{E}^{(m)}\Big(\sum_{s=1}^{\tau}\beta^s(h_{s+t}^{(m)}-\gamma)\Big|\mathcal{F}_t^{(m)}\Big)\leq 0$$

where $\mathcal{T}^{(m)}=\{(\mathcal{F}_{s+t}^{(m)})_{s\geq 0}\text{-positive stopping times}\}$.

Theorem

Define

$$\rho_t^* = \arg\min_m \gamma^{(m)}_{s^{(m)}(t)}$$

where $s^{(m)}(t)$ *indicates the number of plays of bandit* m *by time* t.

Then ρ^* *is a dynamic Gittins' optimum for the discounted cost* $g_t^\rho = \beta^t h^{(\rho_t)}_{s^{(m)}(t)}$.

NB. As $\gamma^{(m)}$ only changes when machine m is played, this is a consistent strategy: at time t we plan to play until $\gamma^{(m)}$ is no longer minimal, and this is our eventual strategy.

- The purpose of this theorem is that we can compute $\gamma_s^{(m)}$ by solving one inverse problem based on optimal stopping, rather than an m-dimensional problem
- The result does not make any assumptions on the costs, information flows or stationarity of any bandit (apart from independence, and they only can change when you play them).
- Robust optimal stopping, with recursive nonlinear expectations, has been well studied, and each problem is of this type.
- If our machines are Markovian, then we can solve this via a pde/difference equation method. If all machines are symmetric, then only one problem needs to be solved.
- We can identify qualitative phenomena from this solution.

- Now that we have our optimality result, we can start analysing our problem numerically
- We will consider a simple setting, where our machines return simple Bernoulli costs of 0 or 1, with a constant but unknown probability.
- We can solve the robust optimal stopping problem, for any proposed γ, using backward recursion.
- We use a horizon of $T^{(m)} = 10000$ and have a discount rate $\beta \approx 1$.

- At each time, we have an estimated probability $p = \mathbb{P}(h = 1) \approx \hat{p}_t$ which evolves like a Bayesian estimate, with implied initial sample size $T_0 = 1$.
- We infer credible intervals Θ_t using the conjugate Beta prior, with a fixed credibility level k
- We define our nonlinear expectation by

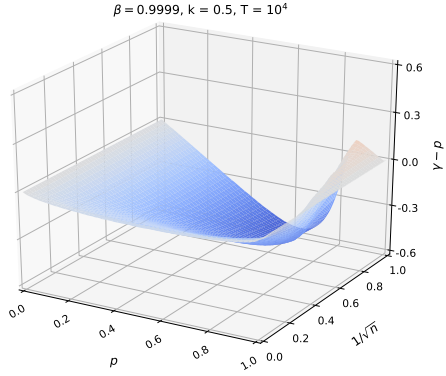
$$\mathcal{E}_t^{(m)}(\xi) = \sup_{p \in \Theta_t} \mathbb{E}_p[\xi | \mathcal{F}_t^{(m)}]$$

  for $\mathcal{F}_{t+1}^{(m)}$-mle ξ, and then extend by backward recursion.
- A similar framework is in Bielecki, Cialenco & Chen.

Low uncertainty aversion

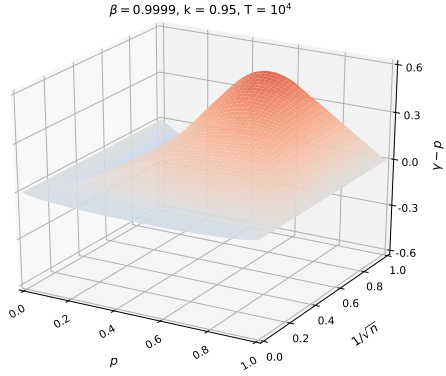
Exploring the unknown requires tolerating uncertainty. —Brian Greene

Mathematical Institute



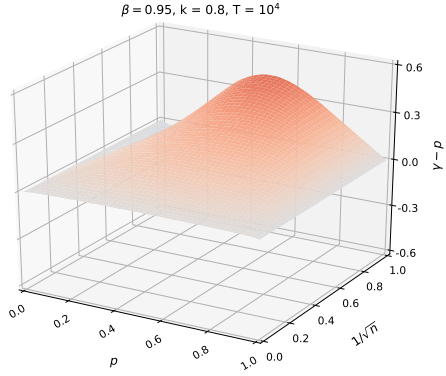
High uncertainty aversion

Exploring the unknown requires tolerating uncertainty. —Brian Greene



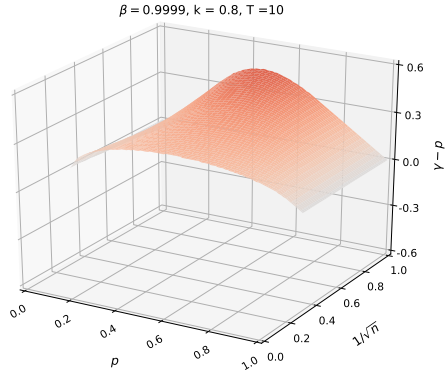
Strong discounting

Exploring the unknown requires tolerating uncertainty. —Brian Greene



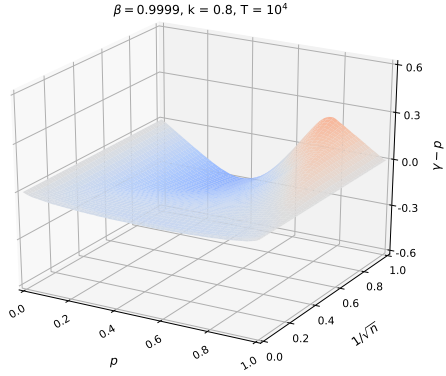
Short horizon

Exploring the unknown requires tolerating uncertainty. —Brian Greene



A ‘balanced’ setup

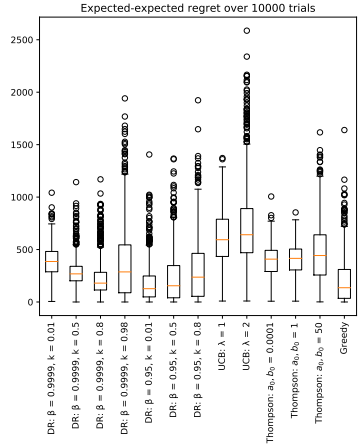
The real problem is not whether machines think but whether men do. —B.F. Skinner

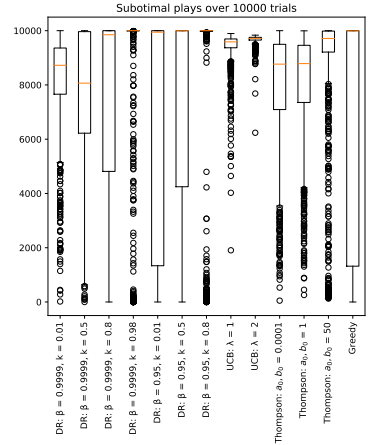


- We can also simulate behaviour for a particular setup.
- Consider $M = 50$ machines, with varying (true) probabilities of a cost.
  - For each trial, we simulate a, b from $\Gamma(1, 1/100)$, then each bandit's probability from $\text{Beta}(a, b)$.
- We can compute performance of various expectations of this type, both in terms of regret (average performance vs. perfect play), and in terms of how often the better arm is chosen.

Behaviour over $T = 10000$ steps

An optimist is a guy that has never had much experience. —Don Marquis





Conclusions

An optimist is a guy that has never had much experience. —Don Marquis

- Further work on the numerical side is needed
  - A continuous version and asymptotics may be interesting
  - More delicate uncertainty frameworks may be important for performance
- Understanding the connection between learning, uncertainty and probability distortions is of interest
- Novel optimality criteria can address dynamic inconsistency in useful ways
- Ultimately, gaining any insight into the not-independent bandits case would be very interesting in practice