Regret learners and bounded rational inductive agents

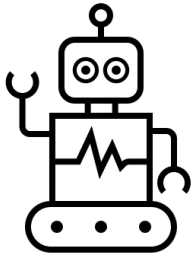
Caspar Oesterheld

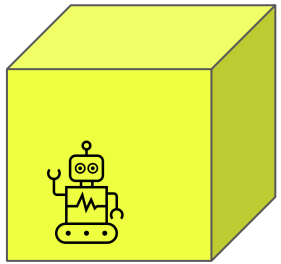
The plan

- The problem setting
  - ~“adversarial” multi-armed bandit problems
  - Issues:
    - Bounded rationality
    - Paradoxes of self-references
    - Newcomb-like problems
    - Counterfactuals
- Two or three approaches
  - Regret learning
  - (Optional: Predictors (e.g., Garrabrant inductors) + importance-weighted estimation)
  - Bounded rational inductive agents

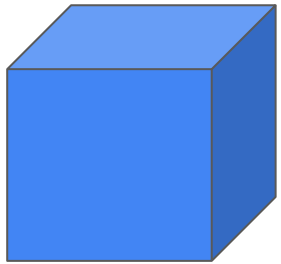
The problem setting

Multi-armed bandits

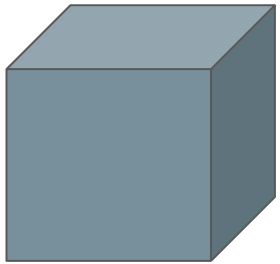




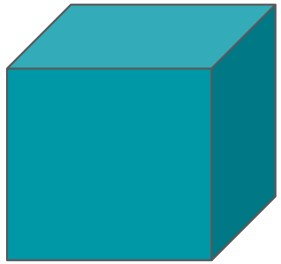
Reward: 3

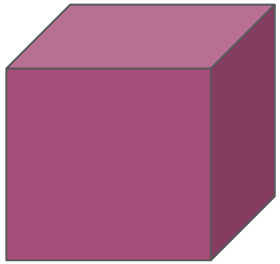


Reward: ?



Reward: ?





Goal: Learn to choose the (myopically) best box.

Meta-goal: Specify what it means to be a rational learner in this setting.

A conceptually easy example – learning the means

In the beginning, environment generates $\mu_1,...,\mu_{10}$~Uniform([0,1]).

Reward of box i ~ Normal(μ_i,1)

A rational agent should, given enough time, figure out what the best box is and in the limit learn to pull that arm with frequency 1.

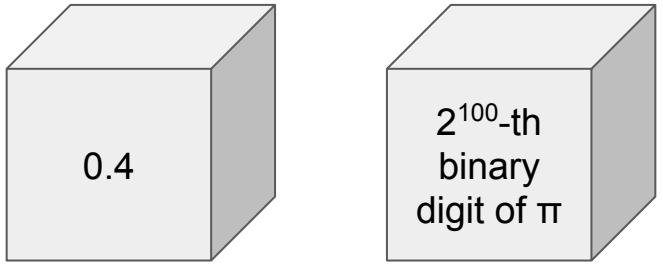
(To guarantee optimality with probability 1, you need to take each box (1,...,10) infinitely many times.)

How do you learn the optimal arm fastest?

- For the purpose of this talk (and most of my work), we don't really care.
- But there's lots of work on this sort of question in the multi-armed bandit literature.

Cases of bounded rationality

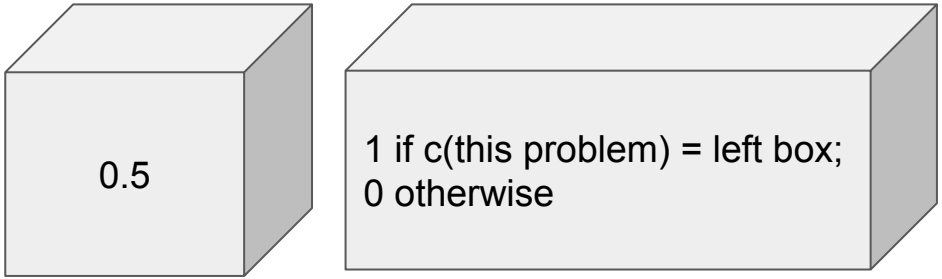
Say you've already learned (mathematical) language…



- In some sense, you should take the right-hand box iff the 2^{100}-th binary digit of π is 1, but that might be difficult…
- Not knowing the 2^{100}-th digit of π, you should take the right box!

Paradoxes of self reference and Newcomb-like scenarios

Agent is computer program c.



- Clearly you shouldn't take the right box.
- If you know that you're taking the left box, you should take the right box.

General idea: consider rationality relative to hypothesis class

General idea: consider rationality relative to hypothesis class

- Consider hypothesis set.
  - E.g.:
    - All polynomial-time algorithms (bounded rationality)
    - Your friends Alice, Bob and Charlie who advise you
  - Generally contains some bad / false hypotheses.
- Rationality requirement of the form:
  "The agent shouldn't be outperformed by any of the given hypotheses (in the limit)."
- The crux is: What is the "no outperformance" condition to satisfy?
- Examples:
  - Garrabrant inductors
  - In this talk:
    - Regret learning
    - Bounded rational inductive agents
  - Bayes? Infra-Bayes?
  - Negative example: Train a single neural net with gradient descent.
- For simplicity: Assume a *finite* set of hypotheses for this talk.
  (Extending to the infinite case is a matter of accounting.)
- People like to give the hypotheses different names: traders, experts, bidders, …

Regret learning – preliminaries

A *bandit problem* specifies the following at each time t:

- a finite list of available boxes (“arms”),
- a function R_t mapping boxes to rewards.

These may depend on agent choices at times 1,...,t-1. (“Reactive” bandit)

An *expert* (for the given bandit problem) specifies at each time t one of the available boxes.

Regret learning

At each time t, agent chooses box B_t to get reward $R_t(B_t)$.

Consider a specific expert who recommends boxes B_t' which obtain (counterfactual) rewards $R_t(B_t')$.

Then the cumulative regret at time T to the expert is

$$\Sigma_{t=1,..,T} R_t(B_t') - R_t(B_t).$$

Rationality criterion: A rational (*Hannan-consistent* / *no regret*) learner for a given set of experts E has sublinear regret (as $T \to \infty$) to each expert in E and in all bandit problems.

(Sublinear regret $\Leftrightarrow$ Average per-round regret$\to 0$ as $T \to \infty$.)

For an introduction, see: Lattimore and Szepesvári: Bandit Algorithms. Available for free online.

Compare ratificationism in the decision theory of Newcomb-like problems.

Regret learning: conceptually simple example

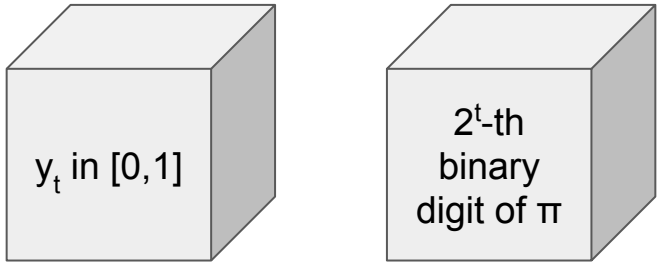
In the beginning, environment generates $\mu_1,...,\mu_{10}$~Uniform([0,1]).

Reward of box i ~ Normal(μ_i,1)

Assume you have at least the 10 constant experts.

Sublinear regret ⇔ take $\text{argmax}_i\ \mu_i$ with limit frequency 1.

Regret learning and bounded rationality



Say you have the following three experts:

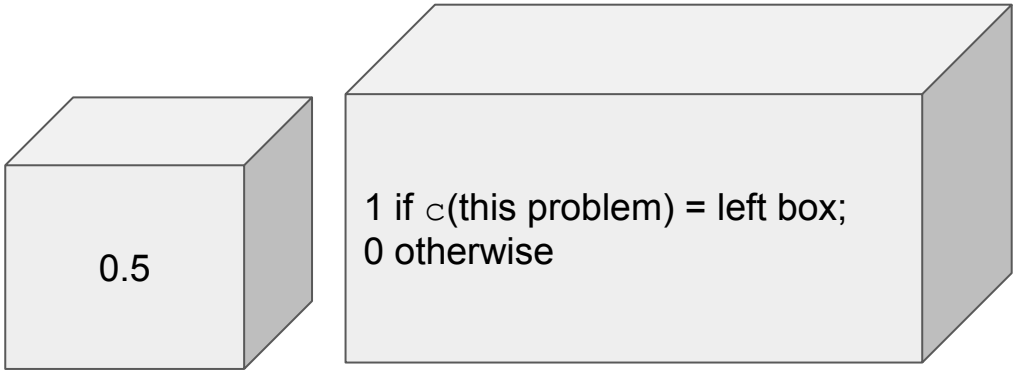
- The one that always recommends left;
- The one that always recommends right; and
- The one that recommends left if $y_t \geq ½$ and right otherwise.

Example agents – do they ensure sublinear regret?

- Following the third expert? ✔
- Following the first or second expert? ✘
- Right box if 2^t-th binary digit of π is 1; left box otherwise. ✔

Deterministic agents can’t guarantee sublinear regret

Agent is `computer program c.`

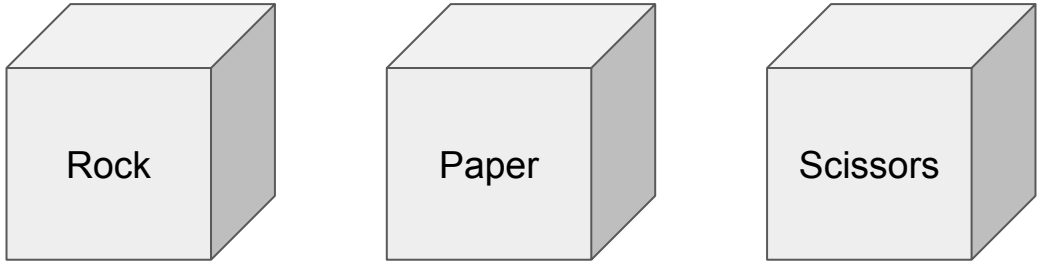


- If you pick the left box with limit frequency 1, you have linear regret to the expert who always recommends the right box.
- Otherwise you have regret to the expert who always recommends the left box.

Solution in regret learning: use randomization

- Assumption: The agent can independently (from the environment) randomize.
- Experts may be stochastic, but are not allowed to recommend probability distributions!

Example: Rock–paper–scissors against Omega



Imagine expert A samples its recommendation Rock/Paper/Scissors uniformly at random.

Then low regret requires convergence to mixing uniformly over Rock, Paper, Scissors.

Existence of regret minimizers

Even with independent randomization, it is not obvious whether we can design an algorithm that ensures sublinear regret… However, it turns out:

Theorem: For any (finite) set of experts, there is an algorithm that takes expert advice as input, chooses boxes stochastically and guarantees sublinear regret w.p. 1 for all (potentially “reactive”) bandit problems.

This result seems conceptual. One might hope that it has a conceptual proof?

Alas…

The algorithm*

Extremely high-level:

Keep track of how well each expert does when following their recommendation.

Randomize over what expert to follow, giving higher probability mass to experts that did well.

Input: n, k, η

Set $\hat{S}_{0i} = 0$ for all i

for $t = 1, \ldots, n$ **do**

Calculate the sampling distribution P_t:

$$P_{ti} = \frac{\exp\left(\eta \hat{S}_{t-1,i}\right)}{\sum_{j=1}^{k} \exp\left(\eta \hat{S}_{t-1,j}\right)}$$

Sample $A_t \sim P_t$ and observe reward X_t

Calculate $\hat{S}_{ti}$:

$$\hat{S}_{ti} = \hat{S}_{t-1,i} + 1 - \frac{\mathbb{I}\{A_t = i\}(1 - X_t)}{P_{ti}}$$

end for

Algorithm 9: Exp3.

$$\eta = \sqrt{\log(k)/(nk)}$$

Screenshots from Lattimore and Szepesvári: Bandit Algorithms. Available for free online.

*Actually, this is the algorithm for a slightly simpler problem.

The proof I

Proof For any arm i, define

$$R_{ni} = \sum_{t=1}^{n} x_{ti} - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right],$$

which is the expected regret relative to using action i in all the rounds. The result will follow by bounding R_{ni} for all i, including the optimal arm. For the remainder of the proof, let i be some fixed arm. By the unbiasedness property of the importance-weighted estimator $\hat{X}_{ti}$,

$$\mathbb{E}[\hat{S}_{ni}] = \sum_{t=1}^{n} x_{ti} \qquad \text{and also} \qquad \mathbb{E}_{t-1}[X_t] = \sum_{i=1}^{k} P_{ti} x_{ti} = \sum_{i=1}^{k} P_{ti} \mathbb{E}_{t-1}[\hat{X}_{ti}]\,. \tag{11.8}$$

The tower rule says that for any random variable X, $\mathbb{E}[\mathbb{E}_{t-1}[X]] = \mathbb{E}[X]$, which together with the linearity of expectation and Eq. (11.8) means that

$$R_{ni} = \mathbb{E}\left[\hat{S}_{ni}\right] - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{k} P_{ti}\hat{X}_{ti}\right] = \mathbb{E}\left[\hat{S}_{ni} - \hat{S}_n\right], \tag{11.9}$$

The proof II

where the last equality serves as the definition of $\hat{S}_n = \sum_t \sum_i P_{ti} \hat{X}_{ti}$. To bound the right-hand side of Eq. (11.9), let

$$W_t = \sum_{j=1}^{k} \exp\left(\eta \hat{S}_{tj}\right) .$$

By convention an empty sum is zero, which means that $\hat{S}_{0j} = 0$ and $W_0 = k$. Then,

$$\exp(\eta \hat{S}_{ni}) \leq \sum_{j=1}^{k} \exp(\eta \hat{S}_{nj}) = W_n = W_0 \frac{W_1}{W_0} \cdots \frac{W_n}{W_{n-1}} = k \prod_{t=1}^{n} \frac{W_t}{W_{t-1}} . \quad (11.10)$$

The ratio in the product can be rewritten in terms of P_t by

$$\frac{W_t}{W_{t-1}} = \sum_{j=1}^{k} \frac{\exp(\eta \hat{S}_{t-1,j})}{W_{t-1}} \exp(\eta \hat{X}_{tj}) = \sum_{j=1}^{k} P_{tj} \exp(\eta \hat{X}_{tj}) . \quad (11.11)$$

We need the following facts:

$$\exp(x) \leq 1 + x + x^2 \text{ for all } x \leq 1 \quad \text{and} \quad 1 + x \leq \exp(x) \text{ for all } x \in \mathbb{R} .$$

The proof III

Using these two inequalities leads to

$$\begin{aligned}\frac{W_t}{W_{t-1}} &\leq 1 + \eta \sum_{j=1}^{k} P_{tj} \hat{X}_{tj} + \eta^2 \sum_{j=1}^{k} P_{tj} \hat{X}_{tj}^2 \\ &\leq \exp\left(\eta \sum_{j=1}^{k} P_{tj} \hat{X}_{tj} + \eta^2 \sum_{j=1}^{k} P_{tj} \hat{X}_{tj}^2 \right) . \end{aligned} \tag{11.12}$$

Notice that this was only possible because $\hat{X}_{tj}$ is defined by Eq. (11.6), which ensures that $\hat{X}_{tj} \leq 1$ and would not have been true had we used Eq. (11.3). Combining Eq. (11.12) and Eq. (11.10),

$$\exp\left(\eta \hat{S}_{ni}\right) \leq k \exp\left(\eta \hat{S}_n + \eta^2 \sum_{t=1}^{n} \sum_{j=1}^{k} P_{tj} \hat{X}_{tj}^2 \right) .$$

Taking the logarithm of both sides, dividing by $\eta > 0$ and reordering gives

$$\hat{S}_{ni} - \hat{S}_n \leq \frac{\log(k)}{\eta} + \eta \sum_{t=1}^{n} \sum_{j=1}^{k} P_{tj} \hat{X}_{tj}^2 . \tag{11.13}$$

As noted earlier, the expectation of the left-hand side is R_{ni}. The first term on the right-hand side is a constant, which leaves us to bound the expectation of the second term. Letting $y_{tj} = 1 - x_{tj}$ and $Y_t = 1 - X_t$ and expanding the definition

The proof IV

of $\hat{X}_{tj}^2$ leads to

$$\begin{aligned}
\mathbb{E}\left[\sum_{j=1}^{k} P_{tj}\hat{X}_{tj}^2\right] &= \mathbb{E}\left[\sum_{j=1}^{k} P_{tj}\left(1-\frac{\mathbb{I}\{A_t=j\}\,y_{tj}}{P_{tj}}\right)^2\right] \\
&= \mathbb{E}\left[\sum_{j=1}^{k} P_{tj}\left(1-2\frac{\mathbb{I}\{A_t=j\}\,y_{tj}}{P_{tj}}+\frac{\mathbb{I}\{A_t=j\}\,y_{tj}^2}{P_{tj}^2}\right)\right] \\
&= \mathbb{E}\left[1-2Y_t+\mathbb{E}_{t-1}\left[\sum_{j=1}^{k}\frac{\mathbb{I}\{A_t=j\}\,y_{tj}^2}{P_{tj}}\right]\right] \\
&= \mathbb{E}\left[1-2Y_t+\sum_{j=1}^{k} y_{tj}^2\right] \\
&= \mathbb{E}\left[(1-Y_t)^2+\sum_{j\neq A_t} y_{tj}^2\right] \\
&\leq k\,.
\end{aligned}$$

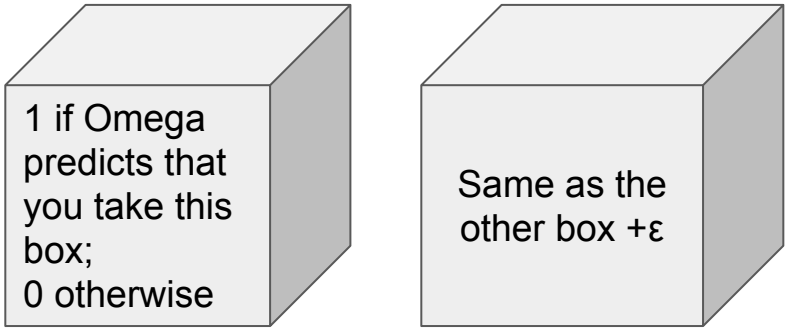
Summing over t, and then substituting into Eq. (11.13), we get

$$R_{ni} \leq \frac{\log(k)}{\eta}+\eta nk = 2\sqrt{nk\log(k)}\,,$$

where the equality follows by substituting $\eta=\sqrt{\log(k)/(nk)}$, which was chosen to optimise this bound. □

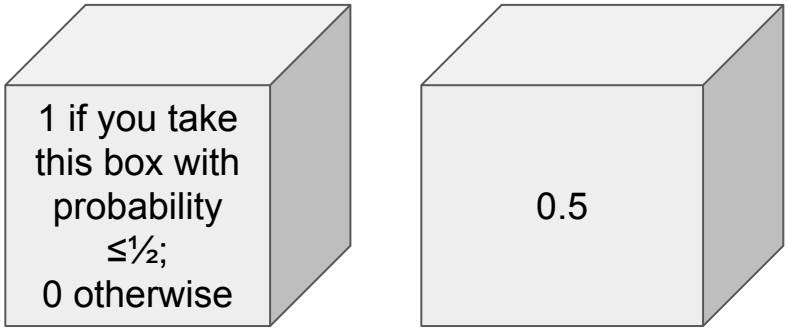
Why I find regret learning unsatisfactory

Newcomb’s problem



- Optimal (EDT/TDT/UDT/FDT/…): Take the left box!
  (assuming Omega is a good enough predictor)
- But taking the left box incurs a regret of epsilon (to the “always the right box” expert).
- Regret learning requires that you learn to always take the right box.

Regret learning is a waste of randomizing



- Ideally take the left box with probability ½.
- But this incurs linear regret (to the expert who says you should always take the left box).
- Depending on your set of experts, regret learning requires some wacky oscillation…
- This all holds even if one of the experts tells you exactly what's going on.

Exploration?

- You'll often hear people refer to randomization in regret learning as *exploration*.
- They're not telling you the whole truth!
- Randomization is also about:
  - Making some experts do worse (to achieve sublinear regret).
  - Making the counterfactuals well defined.

Optional: Prediction + importance-weighted estimation

(Importance-weighted estimators are one ingredient of regret learning algorithms.)

Idea 1: Don’t we want to maximize expected utility?

Imagine we could somehow get something that predicts expected values.

Then shouldn’t we just take the box B that maximizes **E**[R(B)]?

Idea 2: We could learn to estimate expected utilities

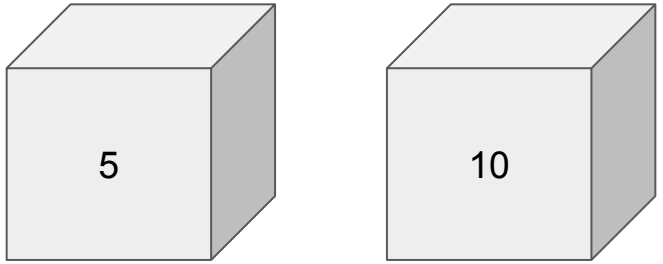
Say in each time step t, we have taken box B_t and received reward r_t.

Then could, e.g., use squared error loss to train a model on training data set

$\{B_t \rightarrow r_t\}_t$.

(Or use something like a Garrabrant inductor, etc.)

Problem: Counterfactuals



Model: E[5] = 5 E[10] = –100

- Following the model, the agent always takes the left box.
- The incorrect prediction about the right box is never refuted.

Compare:
Othman and Sandholm (2010): Decision Rules and Decision Markets. AAMAS.
Garrabrant (2017): Two Major Obstacles for Logical Inductor Decision Theory. Alignment Forum.
“5 and 10 problem”

Idea 3: randomized choice + importance-weighted estimation

At each time t choose probability distribution σ_t over the boxes.

May give most weight to the optimal action, but have to give positive (not too quickly vanishing) probability to all actions.

Then from each time step t we get the following data:

- $B_t \rightarrow R(B_t)/\sigma_t(B_t)$ for B_t that was in fact chosen.
- $B_t \rightarrow 0$ for all B_t not taken.

Why? Because:

$E[\mathbf{1}[B_t \text{ chosen}] * R(B_t)/\sigma_t(B_t)] = E[R(B_t)]$.

$\mathbf{1}[B_t \text{ chosen}] * R(B_t)/\sigma_t(B_t)$ is called an importance-weighted estimator of $R(B_t)$.

Training to predict the mean of $R(B_t)$
~ training to predict mean of $\mathbf{1}[B_t \text{ chosen}] * R(B_t)/\sigma_t(B_t)$.

Cf. Yiling Chen, Ian A. Kash, Mike Ruberry, and Victor Shnayder (2014): Eliciting Predictions and Recommendations for Decision Making. ACM TEAC.

Bounded rational inductive agents

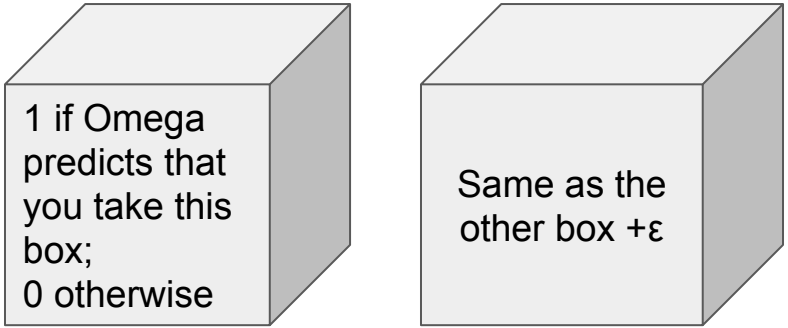
Oesterheld, Demski, Conitzer (2023): A theory of bounded inductive rationality. TARK ‘23.

The algorithm and the criterion

Regret learning:	algorithm	**criterion**
Importance-weighted estimation:	**algorithm**	criterion
(Garrabrant inductors:)	**algorithm**	criterion
Bounded rational inductive agents:	**algorithm**	criterion

Hypothesis type

We consider hypotheses (formerly experts, now sometimes also *bidders*) that recommend (as usual) and estimate.



Intuition: Bidder says, “I think you should take the left box, and I promise that if you do so, you will get an expected reward ≥1.”

Decision auctions

At each time t:

- Run a (first-price) auction between the bidders (a.k.a. hypotheses/experts).
- The winning bidder pays their bid (in “logical dollars”) gets to tell the agent to choose a box.
- The winning bidder receives (in “logical dollars”) the reward received by the agent.

Also: hand out an “allowance” (for exploration).
E.g. (in case of finite set of bidders): each bidder gets 1/t at time t.

High-level intuition for why this works

- Bidders who overbid run out of money.

  ⇒ Agent behavior is controlled by bidders who “keep their promises” (on average).

- If agent behavior is suboptimal, a bidder who knows better can bid higher and recommend the better action.

Conceptually easy problem

In the beginning, environment generates $\mu_1,...,\mu_{10} \sim$ Uniform({0,0.01,0.02,...,0.99,1}).

Reward of box i sampled from ~Normal(μ_i,1).

Say for each i from {1,...,10} and each v from {0,0.01,0.02,...,0.99,1} there is bidder who recommends i and estimates v-ε at every time step.

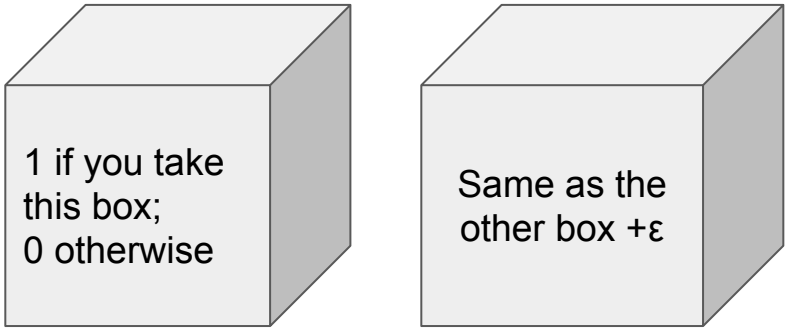
Bidders (i,v) with $v > \mu_i$ lose money whenever they win the auction.

⇒ They play no role in the limit.

Of the remaining bidders, the winners in the auctions will be the ones whose v is highest, i.e., whose μ_i is highest.

Newcomb’s problem

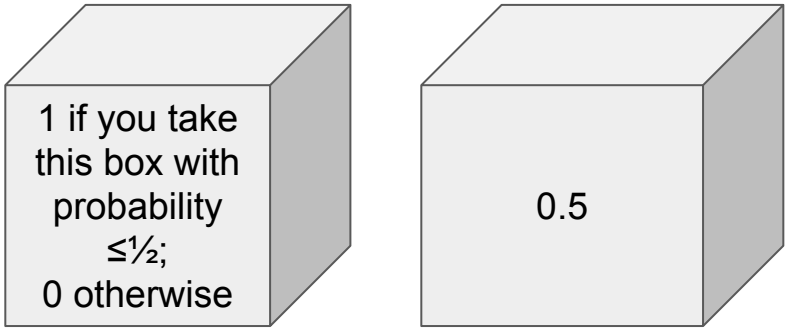
Say you’ve already learned (mathematical) language…



Decision auction learns to take the left box.

No need for randomization!

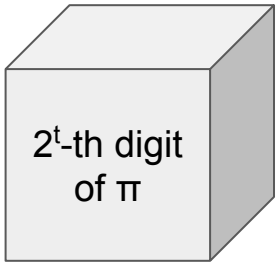
Optimizing over probability distributions



- Consider bidders who recommend the mixed strategy (½,½) and estimate ¾ – ε.
- W.p. 1 these bidders are profitable in the limit.
- Unless another bidder manages to bid ≥ ¾ – ε and “hold its promises”, these bidders win the auction most of the time.

Bounded rational inductive agency – the criterion I

- First part of rationality criterion:
  Agent does not overestimate reward on average in the limit.
- Example: Say the agent chooses at each time step:



- For this requirement:
  Can (presumably) estimate 0.5 or 0.38 all the time; or alternate 0,1.
  Cannot estimate 0.6 all the time.

Bounded rational inductive agency – the criterion II

- **Definition**: If

  hypothesis h's estimate in round t > agent's estimate in round t,

  we say the *agent rejects h in round t*.

- Second part of rationality criterion: If agent rejects h infinitely often, then:
  - h must be tested infinitely often.
  - In the tests of h up until time t, the rewards of h's recommendations "substantially" underperform its estimates.

Thanks!