Philosophical questions about machine learning theory: Online Learning, Bandit Algorithms, and Reinforcement Learning

Maximilian Kasy

March 12, 2021

Some context for this talk

Disclaimer:

- I am interested in conceptual questions,
- but know little about modern philosophy.

My main research interest is **methodology**:

- *Identification:*
  How can we learn from observation about the world?
- *Decision problems:*
  How can we act optimally given partial knowledge about the world?
- *Statistics in a social context:*
  Understanding quantitative methods beyond the framework of single agent decision theory?

Reading machine learning theory

Reading machine learning theory is intriguing in this context:

- Fully automated learning, no human discretion.
- Close to an ideal of methodology:
  Eliminating any dependence of conclusions on the observer's identity.
- Consistently decision-theoretic ("pragmatic"):
  Learning is evaluated based on the loss induced by actions.

Three settings in machine learning

1. Adversarial online learning:
   - Sequential predictions
   - in an unstable, adversarial world.
   - Implications for “induction?”
2. Multi-armed bandits:
   - Sequential (treatment) decisions (interventions),
   - discussed without a language of causality.
   - Implications for “metaphysics of causality?”
3. Reinforcement learning (for games):
   - Sequential moves in a game,
   - treating the adversary as part of the environment.
   - Implications for “solipsism?”

Some references

- **Online learning:**
  Cesa-Bianchi, N. and Lugosi, G. (2006).
  Prediction, learning, and games.
  Cambridge University Press.

- **Multi-armed bandits:**
  Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018).
  A Tutorial on Thompson Sampling.
  Foundations and Trends® in Machine Learning, 11(1):1–96.

- **Reinforcement learning:**
  Sutton, R. S. and Barto, A. G. (2018).
  Reinforcement learning: An introduction.
  MIT press.

Commonalities of these three settings

- Sequential decision making at times $t = 1, 2, \ldots$.
- Decisions result in a loss L_t.
  Good algorithms minimize cumulative loss $\sum_t L_t$.
- Some additional information is revealed at the end of time t.
- Different decisions:
  1. Online learning: Prediction.
  2. Bandit problems: Treatment choice.
  3. Reinforcement learning: Move in a game.

Adversarial online learning

Multi-armed bandits

Reinforcement learning and games

Adversarial online learning

Setup:

- Every period t, we want to make a prediction $\hat{y}_t$ of an outcome y_t.
- Our predictions result in a loss

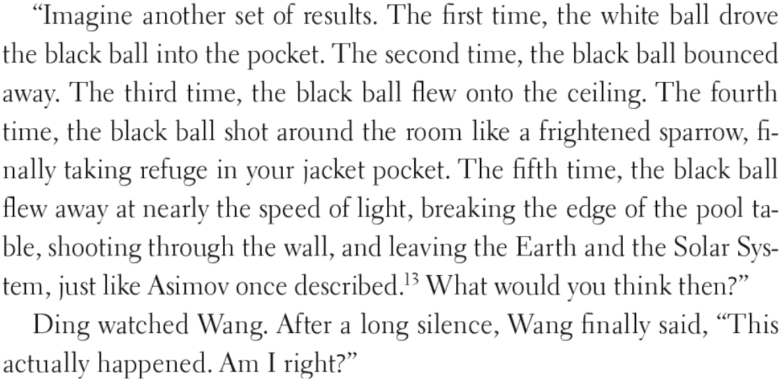
  $$L_t = L(\hat{y}_t, y_t)$$

  that is convex in $\hat{y}_t$, bounded by $[0, 1]$.
- There are $i = 1, \ldots, N$ “experts” (hypotheses, theories, models, model parameters) delivering different predictions $\hat{y}_{i,t}$, resulting in loss

  $$L_{i,t} = L(\hat{y}_{i,t}, y_t).$$

A chaotic, evil world

- **No assumption** is made about how the outcomes y_t are generated.
- We are interested in worst case behavior over all possible sequences $y_1, y_2, \ldots$



"Imagine another set of results. The first time, the white ball drove the black ball into the pocket. The second time, the black ball bounced away. The third time, the black ball flew onto the ceiling. The fourth time, the black ball shot around the room like a frightened sparrow, finally taking refuge in your jacket pocket. The fifth time, the black ball flew away at nearly the speed of light, breaking the edge of the pool table, shooting through the wall, and leaving the Earth and the Solar System, just like Asimov once described.[13] What would you think then?"

Ding watched Wang. After a long silence, Wang finally said, "This actually happened. Am I right?"

Liu Cixin, The Three Body Problem

Worst case regret

- How could we possibly learn anything under such circumstances?
- Adversarial online learning provides an answer.
- Consider regret: How much worse do we do relative to any "expert?"
- Formally: **Average regret**, relative to expert *i*:

$$R_T(i) = \frac{1}{T} \sum_{t=1}^{T} \left[L_t - L_{i,t} \right].$$

- We would like worst case average regret to go to 0 (fast):

$$\max_i R_T(i) \to 0.$$

An algorithm, with regret guarantee

[Slide optional for our discussion]

- **Exponential weighting:** Let

$$w_{i,t} = \exp\left[-\eta \sum_{s=1}^{t} L_{i,t}\right]$$

- Choose the weighted average prediction

$$\hat{y} = \frac{\sum_i w_{i,t} \cdot y_{i,t}}{\sum_i w_{i,t}}.$$

- **Theorem**: For a good choice of η, worst case regret is bounded by

$$\max_i R_T(i) \leq \sqrt{\frac{\log N}{2T}},$$

which vanishes as T gets large.

Discussion

- We can do essentially as well as the best of our experts.
- No matter how the sequence y_t is generated!
- No stability or invariance in the world is assumed.
- A possible way to address the **induction problem**?
- We are guaranteed do well *if anyone can do well*.

Aside:

- This framework doesn't invoke probability anywhere.
- But the proposed predictor also make sense from a Bayesian point of view,
- for a stable stochastic data generating process: "Bayesian model averaging."

Is this good enough?

> *The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.*
>
> Bertrand Russell, The Problems of Philosophy.

- Should our regret bound provide consolation to the chicken?

Adversarial online learning

Multi-armed bandits

Reinforcement learning and games

Multi-armed bandits

Setup:

- Every period t, we want to make a (treatment) decision $d_t \in \{1, \dots, k\}$.
- Our decision results in a **loss**
$$y_t = y_t^{d_t}.$$
- After the period, we **observe** the loss $y_t^{d_t}$ of the action d_t that we chose, but we don't observe the loss y_t^d of the other actions $d \neq d_t$ that we could have chosen.
- Stochastic bandit setup:
  $(y_t^1, \dots, y_t^k)$ is independent and identically distributed over time.
  $\Rightarrow$ Now we do assume **stability** in the world!

Example: Clincal trials

- Patients arrive sequentially.
- Goal: Minimize the number of patients who die.
- We can choose between multiple alternative treatments (including doing nothing).

Exploration versus exploitation

- **Notation**:
  Number of times d was already chosen:
  $$n_t^d = \sum_{s \leq t} \mathbf{1}(d_t = d).$$
  Average loss of d thus far:
  $$\bar{y}_t^d = \frac{1}{n_t^d} \sum_{s \leq t} \mathbf{1}(d_t = d) y_t.$$
- **Exploration:**
  Choose the action that allows you to learn the most.
  E.g., the one with the smallest number of observations n_t^d.
- **Exploitation:**
  Choose the action that you think performs best now.
  E.g., the one with the smallest average loss $\bar{y}_{t-1}^d$.

Upper confidence bound algorithm

- Good algorithms need to **trade off** exploration and exploitation.
- **Upper confidence bound algorithm:**
  Choose the action with the smallest value of
  $$\bar{y}^{d}_{t-1} + B(n^{d}_{t}, t).$$
  for some well-chosen function B that is decreasing in n, increasing in t.
- **"Optimism in the face of uncertainty."**
- **Theorem:** For a good choice of B, average regret (relative to the best action) goes to 0 as fast as $\log(t)/t$.

Discussion

- We could describe the bandit setting in terms of experiments, potential outcomes, treatment effects and causality.
- Strikingly, the literature on bandits largely avoids any talk of causality,
- focusing instead on loss, regret, algorithms.
- Question:
  Do we need a **language / metaphysics of causality**,
  if our goal is just to act successfully?

Adversarial online learning

Multi-armed bandits

Reinforcement learning and games

Reinforcement learning and games

Setup: Markov decision problems

- Like the bandit setup, but additionally there is a state s_t.
- At time t, we observe s_t and then choose an action d_t.
- The action results in loss y_t, and a next period state s_{t+1}.
- Both y_t and s_{t+1} are drawn from stable probability distributions given s_t and d_t.
- The goal is again to minimize the sum of losses $\sum_t y_t$.

Example: Game play

(Action) value functions

- Action value function:
  The expected future sum of losses $\sum_{s \geq t} y_s$,
  given the current state and action.

- One way to get this ("**model based approach**"):
  1. First learn the probability distributions of (s_{t+1}, y_t) given (s_t, d_t).
  2. Then use these to calculate the action value function.

- An alternative way to get this ("**model free approach**"):
  - Directly predict y_t and the next period value function,
  - without attempting to learn a model for the transitions of s_t.

Self-play

The big successes in gameplay were based on this approach:

- Let the computer play against itself.
- Directly **learn the action value function**.
- **Don't attempt to** learn the "model,"
  i.e., **predict the opponent's move**.
- Even though the opponent is just a copy of the same algorithm!
- In competition, pick the move that maximizes the action value.

Discussion

- This is a striking approach:
- **Other players** are treated as **part of the environment**.
- No different from the rules of the game, or the physical environment.
- Even when, in principle, one has a lot of insight into other players: They are exact copies of oneself.
- Connections to **solipsism** and debates about consciousness?

Summary

- We have discussed three dynamic settings in machine learning,
- and connected them to some philosophical questions:
  1. Adversarial online learning
     and the induction problem.
  2. Multi-armed bandits
     and goal-oriented action without a language / metaphysics of causality.
  3. Reinforcement learning and self-play
     and solipsism.

Thank you!