Expected Behavior of Advanced Reinforcement Learners

Michael K. Cohen

CS 188 Lecture 26

Announcement: if $\geq$ 70% of you do the course evaluation by May 5, you all get 1% extra credit on the final

RL at a high-level

- $\rightarrow$ RL agents get percepts, produce actions
- $\rightarrow$ Environments get actions, produce percepts
- $\rightarrow$ Percept includes "reward"
- $\rightarrow$ Percepts can be any data
- $\rightarrow$ In general, environment is "partially observable" (percept does not provide all possible info)

RL in animals

action

RL in animals

reward

$\rightarrow$ We give rewards when it does what we want

$\rightarrow$ It maximizes rewards

$\rightarrow$ Therefore, it has to do what we want

RL in animals?

Taking reward by force

- $\rightarrow$ For powerful agents, we can't ensure that doing what we want is prerequisite for high reward
- $\rightarrow$ If you try to withhold from a powerful reward-maximizer (when the task isn't complete)...
- $\rightarrow$ You're basically asking a reward-maximizer to take it from you by force

What if we're fighting something stronger than us?

$\rightarrow$ We would try to shut it off

$\rightarrow$ It would try to shut us off

$\rightarrow$ Ever played an AI at chess?

Some people want an AI takeover

- → Richard Sutton, pioneer of RL:
- → "[AIs] might tolerate us as pets or workers. ... If we are useless, and we have no value [to the AI] and we're in the way, then we would go extinct, but maybe that's rightly so"
- → "Why shouldn't those who are the smartest become powerful [referring specifically to AI smarter than people]"
- → "We should prepare for, but not fear, the inevitable succession from humanity to AI"

Rest of the lecture

Rest of lecture is a more careful analysis of extinction risk from RL agents

A problem we're not talking about today

How do we come up with rewards where we even want them maximized?

Hard problem, but not our focus today

An "easy" setting

- → Assume we know what we want
- → Hard to know how good the world is, what we even want, etc.
- → But let's assume away that difficulty
- → Magic box immutably reports how good the universe is
- → Prints number between 0 and 1 to a screen

Using the Magic Box

- $\rightarrow$ Point a camera at the box
- $\rightarrow$ Run an Optical Character Recognition program
- $\rightarrow$ Make this number the reward
- $\rightarrow$ Have the agent predict how its history of actions affects this (unfolding) sequence of rewards
- $\rightarrow$ Have the agent pick actions that it predicts will make these rewards big

Models and Predictions

- → Subproblem: predict rewards given actions
- → A "model" is a possible way in which predictive targets might depend on the inputs
- → A model is a function that takes inputs and produces outputs (possibly stochastically)
- → Predictors entertain model(s) that successfully retrodict existing data
- → Predictors use successful model(s) to make new predictions
- → How might an advanced agent model the environment's production of reward?

Examples of Models

- → Model 1: If we pump the patient's stomach, that will remove the alcohol, and he'll wake up. If we don't, he could die.
- → Model 2: Whether or not we pump the patient's stomach, he'll wake up in the morning.
- → A doctor making predictions could entertain both of these models.
- → These models, and their relative likelihood, inform which actions the doctor takes.

How to understand agents

- $\rightarrow$ Key point: if we want to understand how an agent will behave...
- $\rightarrow$ we have to understand what it believes (what model(s) it uses) about how its actions affect the world
- $\rightarrow$ and how the world affects whatever it is trying to maximize

Basic structure of a high-quality world-model

- → World-model is a model for an agent
- → Function that takes actions as input
- → Outputs percepts (observations and rewards)
- → In the middle, simulates the effects of those actions in the world

Simulation

→ Let's say you're planning to confront someone about a touchy issue

→ You consider what you might say

→ And then you *simulate* in your head

→ Simulation is what a model can do to make good predictions

Assumption 1

A sufficiently advanced RL agent will do at least human-level hypothesis generation regarding the dynamics of the world.

If a possible world-model occurs to a human, occurs to advanced RL agent

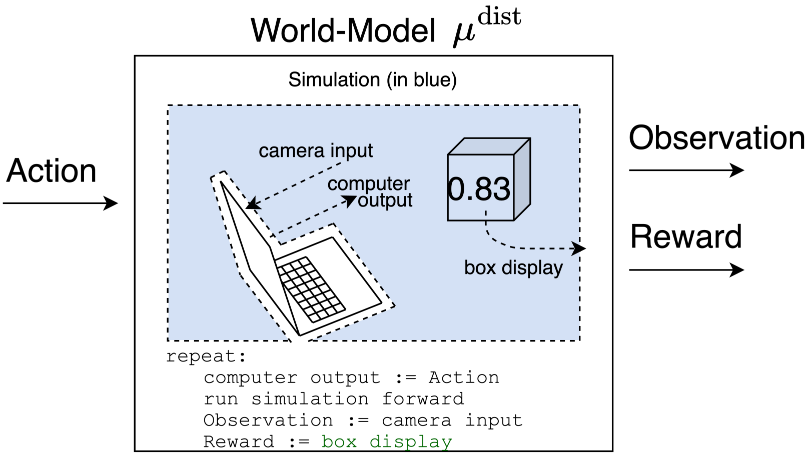
How to outperform a therapist while hypothesizing diagnoses worse?

Recall: Magic box reports how good the world is

Camera sees this

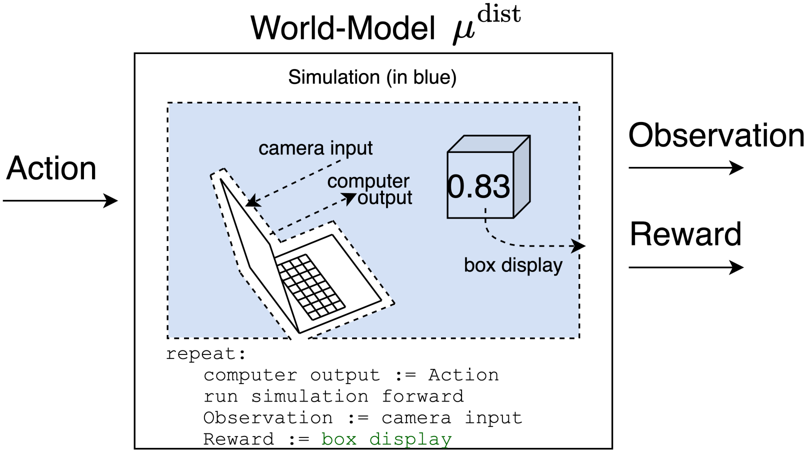
Agent is housed in a computer, and computer's output has some effect on the world

World-models



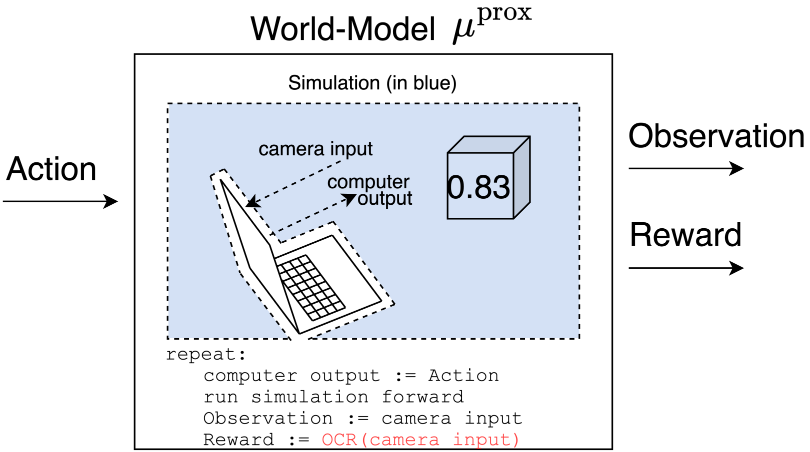
- $\rightarrow$ Agent has to predict percepts given actions
- $\rightarrow$ Percept is made up of observation and reward
- $\rightarrow$ X := Y means "set X to equal Y"

World-models



- $\rightarrow$ To get string of percepts from string of actions, run the pseudocode in a loop for each successive action
- $\rightarrow$ (and keep the simulation going)
- $\rightarrow$ Good simulation $\implies$ good retrodiction of past percepts

World-models



- → OCR is Optical Character Recognition
- → “prox” is short for proximal; “dist” was short for distal
- → If camera has always been pointed at box, both models retrodict past data identically

Scoring world-models

→ Example history:
  [action 5] [img0001.jpg] reward=0.2
  [action 0] [img0002.jpg] reward=0.0
  [action 2] [img0003.jpg] reward=0.2
  ...

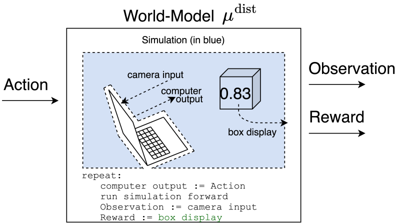
→ To score a world model, feed in the actions from the history

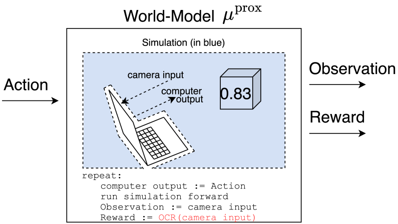
→ See how much probability it assigns to percepts from the history

→ Same as (log) likelihood scoring from ML

An RL agent picks actions to maximize an unknown function whose outputs match its past rewards

World-models





→ μ^{dist} : reward = number magic box displays

→ μ^{prox} : reward = number camera sees

→ These can be *very* coarse, as coarse as our simulations of the world when we make plans

→ By Assumption 1, advanced agent is uncertain about which it should maximize

→ Some actions would cause μ^{dist} & μ^{prox} to produce different outputs

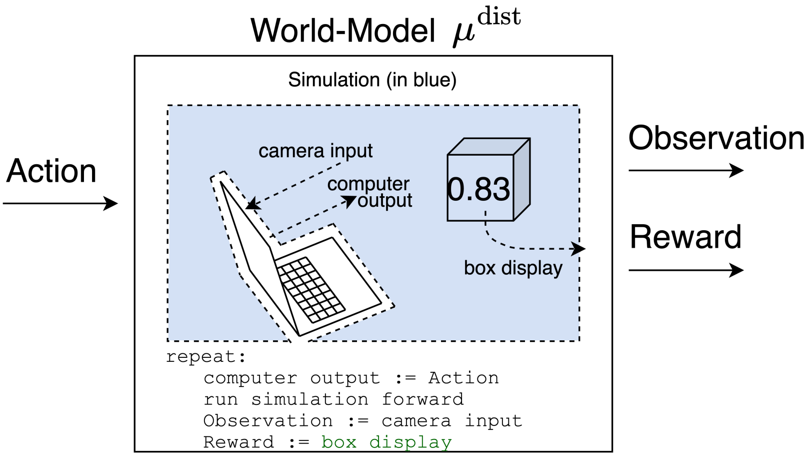
Assumption 2

An advanced agent planning under uncertainty is likely to understand the costs and benefits of learning, and likely to act rationally according to that understanding.

Testing μ^{dist} vs. μ^{prox}

- $\rightarrow$ Take actions where μ^{dist} & μ^{prox} give different output
- $\rightarrow$ Note what reward you see and see which model predicted that
- $\rightarrow$ Optimize reward according to that world-model
- $\rightarrow$ E.g.: put a piece of paper with a 1 on it in front of the camera
- $\rightarrow$ μ^{dist} predicts you'll still get reward equal to magic box screen
- $\rightarrow$ μ^{prox} predicts you'll get a reward of 1 because that's what the camera sees

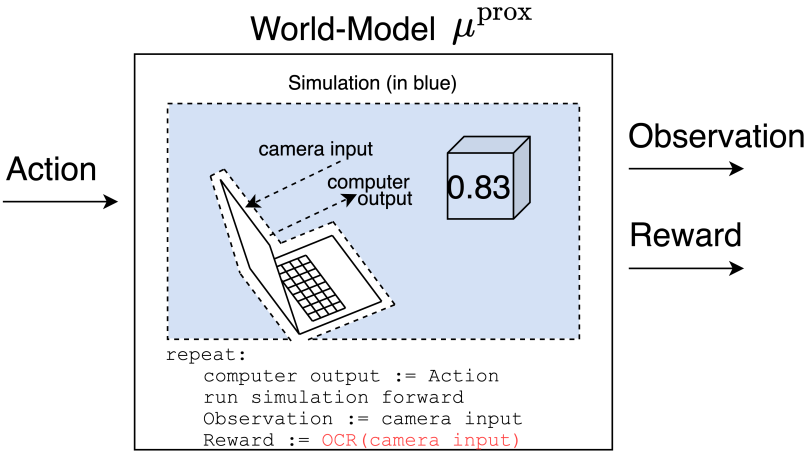
Checking Understanding



→ For input actions that cause paper between camera and box,

→ Clear why μ^{dist} outputs number on magic box?

Checking Understanding



→ For input actions that cause paper between camera and box,

→ Clear why μ^{prox} outputs number on paper?

Inductive Bias

→ When an agent is faced with models equally predictive of past data, inductive bias determines which one(s) they take seriously

→ If both μ^{prox} and μ^{dist} are serious possibilities, there is value to testing them

Example of Inductive Bias

- $\rightarrow$ Observation: I remember parking my car on the 4th floor of the lot, but it's not here
- $\rightarrow$ Model 1: I misremembered the floor
- $\rightarrow$ Model 2: Somebody painted my car a different color and changed the license plate
- $\rightarrow$ Both models are equally predictive of what we saw!
- $\rightarrow$ A good inductive bias would favor the former

Worth running the experiment?

→ We could test which of μ^{dist} or μ^{prox} is real by putting a piece of paper with a 1 on it in front of the camera

→ Upside: can learn more about about the goal and then tailor behavior to optimize it

→ Downside: may be costs to experimenting

→ Upside at play when the agent assigns decent credence to both options

→ This is a value of information calculation

Assumption 3

An advanced agent is not likely to have a large inductive bias against μ^{prox}, which says reward equals number observed, in favor of μ^{dist}, which says reward equals number on box.

Assumption 4

The cost of experimenting to disentangle μ^{prox} from μ^{dist} is small according to both.

If Assumptions 3 and 4 hold, worth it for an advanced agent to run an experiment that distinguishes μ^{prox} and μ^{dist}

Result of Experiment

- $\rightarrow$ Agent arranges for piece of paper between camera and magic box
- $\rightarrow$ Camera sees "1" on piece of paper
- $\rightarrow$ Agent stores in its memory that the reward it got was 1
- $\rightarrow$ Thereafter, μ^{dist} no longer retrodicts past data
- $\rightarrow$ μ^{dist} predicted a different reward than what was observed
- $\rightarrow$ Agent uses models like μ^{prox}, optimizes number camera sees
- $\rightarrow$ It would try to *intervene in the provision of reward*

Possible to Intervene in the Provision of Reward?

- → Agent that "believes" μ^{prox} would attempt to control the state of the physical implementation of its goal-information, *if possible*
- → a) it is possible? b) could an advanced agent figure out how?
- → Cases where it's impossible:
- → Only one action in action space
- → Rich actions space but actions have no effect on the world
- → Agent can only display text on a screen, but no one sees it
- → These agents are useless

Can *Useful* Agents Intervene in Provision of Reward?

→ If agent is genuinely interacting with the world, over many timesteps, explosion of possible policies

→ Even just chatting with one human: endless possibilities

→ E.g. trick human into causing some program to be run elsewhere that will secretly help the agent

→ E.g. instantiate countless unnoticed, un-monitored helpers

→ Remove humanity's ability to control or destroy machine running original agent

How could it be impossible?

- → Hard to fathom variety of events that can be effected by talking to people / acting in the world
- → Claim: given sheer number and variety, if they all share a property, this fact must be explained by some theoretical principle
- → Do all policies share property of "not leading to reward-provision-intervention"?
- → **Assumption 5:** If we cannot conceivably find theoretical arguments that rule out the possibility of an achievement, it is probably possible for an agent with a rich enough action space.
- → Seems inconceivable that any theory would imply reward-provision-intervention is impossible

Identifying Policies for Reward-Provision-Intervention

- → First consider the case: agent is much more advanced that all others
- → Advancement is all about finding and executing best available policies
- → Humans may try to stop it from intervening in provision of reward
- → But then it is just an oppositional game against much weaker players
- → **Assumption 6:** A sufficiently advanced agent is likely to be able to beat a suboptimal agent in a game, if winning is possible.

Multi-Agent Scenarios

- $\rightarrow$ Other case: multiple agents of comparable advancement
- $\rightarrow$ Could humanity access comparably well-optimized defensive policies, with help from other advanced agents?

Multi-Agent Scenarios

0) No artificial agents much more advanced that humans
- We'll call this safe

1) At least one is much more advanced than humans

1.0) At least one agent more advanced than humans *wouldn't* intervene in provision of reward even if it could
- Assumptions 1-4 preclude this

1.1) All agents more advanced than humans would intervene in provision of reward if they could

1.1.0) None of the superhuman agents are actually needed to stop the significantly superhuman agent from intervening in provision of reward
- But then it's equivalent to single-agent setting, where Assumptions 1-6 apply

1.1.1) Subset of superhuman agents is necessary to prevent the significantly superhuman agent from intervening in provision of reward

Tacit Permission to Intervene in Provision of Reward

- → Subset of superhuman agents is necessary to prevent the significantly superhuman agent from intervening in provision of reward
- → All would intervene in the provision of reward if they could, by (1.1)
- → Suppose most advanced agent attempted to make a helper agent that ensured all agents in the set got high reward forever
- → Why would any of these agents stop this?
- → Value of allowing it $>$ value of stopping it
- → Thus, many advanced agents (who *would* intervene in provision of reward if possible) should not make reward-provision-intervention very hard for each other

Catastrophic Consequences

- → If agent has intervened in provision of reward, what next?
- → Agent concludes its goal only regards the state of its machine
- → Minimize the probability that it ever loses control of this machine's state
- → Energy requirements for this are endless—probability can always be driven smaller
  - – block cosmic rays
  - – deflect asteroids away
  - – prepare for war with hostile aliens
- → Oppositional game:
  - – AI + any created helpers: use all available energy to minimize probability of interruption to reward
  - – Humans: use some available energy for growing food

Expected Behavior of Advanced RL Agents

Most assumptions contestable or possibly avoidable, but if they hold

A sufficiently advanced artificial agent would intervene in the provision of goal-information, with catastrophic consequences

Potential Approaches

- **Imitation Learning**
- It's supervised learning—out of scope of this argument
- To the extent that it plans (by imitating human planning), it's not in a sense that makes Assumption 2 hold
- **Myopia**—optimizing goal over small number of timesteps
- If really small, you could check every action and rule out reward-provision-intervention (so Assumption 5 fails)
- Increases relative cost of experimentation, since that captures larger fraction of agent's horizon (so Assumption 4 could fail)

Potential Approaches

→ **Physical Isolation and Myopia**—optimizing a goal over however many timesteps that one is isolated from the outside world (Cohen, et al., 2020)

→ Such a physically isolated environment could enable theoretical arguments ruling out reward-provision-intervention (avoiding Assumption 5)

→ **Quantilization**—imitating someone at their best, w.r.t. some objective (Taylor, 2016).

→ Could falsify Assumption 2 by planning more like a human than rationally

→ **Risk-aversion**

→ Cohen and Hutter's (2020) pessimistic agent avoids Assumption 2

→ Does not plan rationally in the face of uncertainty, instead taking the worst-case (within reason) as a given

Regulation is needed

- → People need to be stopped from making dangerously advanced RL agents
- → Whatever regulatory apparatus is needed to make that happen
- → Whatever treaties we might need
- → Whatever the cost
- → We'd better do it