Lecture 10: Fast Reinforcement Learning

Emma Brunskill

CS234 Reinforcement Learning

Winter 2026

- With some slides from or derived from David Silver, Examples new

Refresh Your Understanding: Multi-armed Bandits

- Select all that are true:
  1. Algorithms that minimize regret also maximize reward
  2. Up to variations in constants, ignoring δ, UCB selects the arm with $\arg\max_a \hat{Q}_t(a) + f(\sqrt{\frac{1}{N_t(a)}})$ (where $f(x)$ indicates a function of x)
  3. UCB still would likely learn to pull the optimal arm more than other arms if we instead used $\arg\max_a \hat{Q}_t(a) + \sqrt{\frac{1}{\sqrt{N_t(a)}} \log(t/\delta)}$
  4. UCB uses $\arg\max_a \hat{Q}_t(a) + b$ where b is a bonus term. Consider $b = 5$. This will make the algorithm optimistic with respect to the empirical rewards but it may still cause such an algorithm to suffer linear regret.
  5. A k-armed multi-armed bandit is like a single state MDP with k actions
  6. Not Sure

Refresh Your Understanding: Multi-armed Bandits Solution

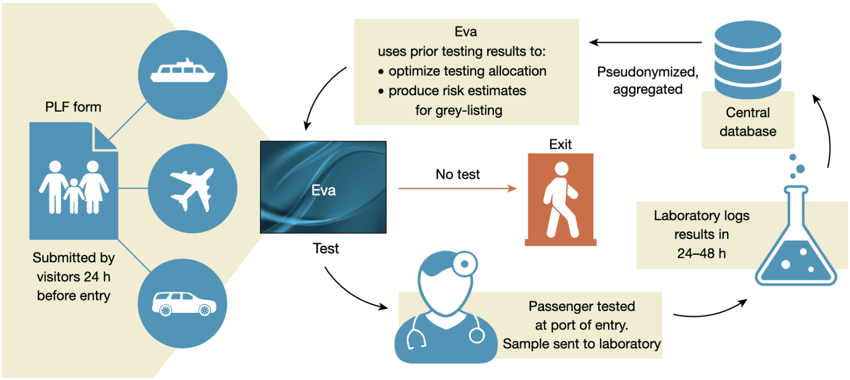
- Select all that are true:
  1. Algorithms that minimize regret also maximize reward
  2. Up to variations in constants, ignoring δ, UCB selects the arm with $\arg\max_a \hat{Q}_t(a) + f(\sqrt{\frac{1}{N_t(a)}})$ (where $f(x)$ indicates a function of x)
  3. UCB still would likely learn to pull the optimal arm more than other arms if we instead used $\arg\max_a \hat{Q}_t(a) + \sqrt{\frac{1}{\sqrt{N_t(a)}} \log(t/\delta)}$
  4. UCB uses $\arg\max_a \hat{Q}_t(a) + b$ where b is a bonus term. Consider $b = 5$. This will make the algorithm optimistic with respect to the empirical rewards but it may still cause such an algorithm to suffer linear regret.
  5. A k-armed multi-armed bandit is like a single state MDP with k actions
  6. Not Sure

  Solutions: True. True. True. True. True

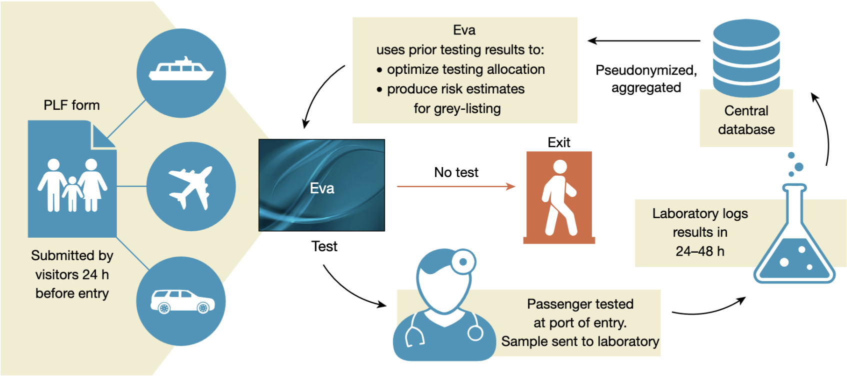
Where We are

- Last time: Bandits and regret and UCB (fast learning)
- This time: Bandits, regret and bayesian bandits (fast learning)
- Next time: Bayesian learning and MDPs (fast learning)

Deciding Who To Test for Covid. Bastani et al. Nature 2001



Deciding Who To Test for Covid. Bastani et al. Nature 2001



- A *nonstationary, contextual, batched bandit problem with delayed feedback and constraints*

Today

- Bandits and regret
- Bandits and Probably Approximately Correct
- Bayesian bandits
- Thompson sampling
- Bayesian Regret

Multiarmed Bandits Notation Recap

- Multi-armed bandit is a tuple of $(\mathcal{A}, \mathcal{R})$
- $\mathcal{A}$: known set of m actions (arms)
- $\mathcal{R}^a(r) = \mathbb{P}[r \mid a]$ is an unknown probability distribution over rewards
- At each step t the agent selects an action $a_t \in \mathcal{A}$
- The environment generates a reward $r_t \sim \mathcal{R}^{a_t}$
- Goal: Maximize cumulative reward $\sum_{\tau=1}^{t} r_\tau$
- **Regret** is the opportunity loss for one step

$$l_t = \mathbb{E}[V^* - Q(a_t)]$$

- **Total Regret** is the total opportunity loss

$$L_t = \mathbb{E}[\sum_{\tau=1}^{t} V^* - Q(a_\tau)]$$

- Maximize cumulative reward $\iff$ minimize total regret

Optimism Under Uncertainty: Upper Confidence Bound (UCB) Algorithm

- UCB1 algorithm

$$a_t = \arg\max_{a \in \mathcal{A}} \left[\hat{Q}(a) + \sqrt{\frac{2 \log \frac{1}{\delta}}{N_t(a)}} \right]$$

Regret Bound for UCB Multi-armed Bandit Sketch (See Section 7.1 of Bandit Algorithms book)

The following proof follows the proof of Theorem 7.1 in the Bandit Algorithms textbook
`https://tor-lattimore.com/downloads/book/book.pdf`.

Regret Bound for UCB Multi-armed Bandit Sketch New

Thm 7.1. Consider UCB on a stochastic K-armed bandit with rewards in 0,1). For any horizon n if $\delta = 1/n^2$ then

$$\mathrm{Regret}_n \leq 3 \sum_{i=1}^{K} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log n}{\Delta_i} \tag{1}$$

Assume w.l.o.g. $a^* = a_1$ and recall:

$$\Delta_i = Q(a^*) - Q(a_i), \qquad \hat{Q} = \text{empirical (no hat = true)} \tag{2}$$

$$\mathrm{Regret}_n = \sum_{i=1}^{K} \Delta_i \, \mathbb{E}[N_t(a_i)] \tag{3}$$

Typo: i —> 1
e.g. UCB_1(t,...)

$$\text{Good event } G_i = \Big\{ \underbrace{Q(a_1) < \min_{t \in n} \mathrm{UCB}_i(t, \delta)}_{*} \Big\} \cap \Big\{ \underbrace{\hat{Q}_{u_i}(a_i) + \sqrt{\frac{2}{u_i} \log \frac{1}{\delta}} < Q(a_1)}_{**} \Big\} \tag{4}$$

$$u_i = \text{some (not yet specified) number of pulls for arm } i \tag{5}$$

Regret Bound for UCB Multi-armed Bandit Sketch New

Δ_i is a fixed (unknown) quantity. We now seek to bound $\mathbb{E}[N_t(a_i)]$:

$$\mathbb{E}[N_t(a_i)] = \mathbb{E}[\mathbf{1}(G_i = T)N_t(a_i)] + \mathbb{E}[\mathbf{1}(G_i^c = T)N_t(a_i)] \tag{6}$$

$$\leq \underbrace{u_i}_{***} + n\,\mathbb{P}(G_i^c = T)(Claim) \tag{7}$$

Proof by contradiction for $\mathbb{E}[\mathbf{1}(G_i = T)N_t(a_i)] \leq u_i$ (***). First assume this is not true. Then there must exist a time step t such that $N_{t-1}(a_i) = u_i$ when arm a_i was pulled/taken:

$$\mathrm{UCB}_i(t-1, \delta) = \hat{Q}_{t-1}(a_i) + \sqrt{\frac{2\log(1/\delta)}{N_{t-1}(a_i)}} \tag{8}$$

$$= \hat{Q}_{t-1}(a_i) + \sqrt{\frac{2\log(1/\delta)}{u_i}} \tag{9}$$

$$< Q(a_i) \quad \text{by good event } (**) \tag{10}$$

$$< \mathrm{UCB}_1(t-1, \delta) \quad \text{by good event } (*) \tag{11}$$

But then we should have selected a_1 instead of a_i. So (***) holds.

Regret Bound for UCB Multi-armed Bandit Sketch New

Next goal: upper bound $\mathbb{P}(G_i^c = T)$. First consider good event (*)

$$
\begin{aligned}
Q(a_1) \geq \min_{t \in n} \mathrm{UCB}_1(t, \delta) &= \bigcup_{s \in n} \left\{ Q(a_1) \geq \hat{Q}_s(a_1) + \sqrt{\frac{2 \log(1/\delta)}{s}} \right\} && (12)\\
\mathbb{P}\Big(Q(a_1) \geq \min_{t \in n} \mathrm{UCB}_1(t, \delta) \Big) &\leq \sum_{s=1}^{n} \mathbb{P}\left(Q(a_1) \geq \hat{Q}_s(a_1) + \sqrt{\frac{2 \log(1/\delta)}{s}} \right) && (13)\\
&\leq n\delta, && (14)
\end{aligned}
$$

where the second line holds from by the union bound, and the third line holds because we assume that each UCB holds with probability δ.

Regret Bound for UCB Multi-armed Bandit Sketch New

Continuing our goal to upper bound $\mathbb{P}(G_i^c = T)$, we next consider the second term in the good event:

$$\left\{ \underbrace{\hat{Q}_{u_i}(a_i) + \sqrt{\frac{2}{u_i}\log\frac{1}{\delta}} < Q(a_1)}_{**} \right\} \tag{15}$$

Assume that we select u_i to be sufficiently big such that, for some $c \in (0,1)$ (to be specified), the following holds

$$\Delta_i - \sqrt{2\log(1/\delta)/u_i} \geq c\Delta_i \tag{16}$$

Then

$$\begin{aligned}
P(\hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_1) &=. \quad P(\hat{\mu}_{i,u_i} + \sqrt{\frac{2\log(1/\delta)}{u_i}} \geq \mu_i + \Delta_i) && (17)\\
&= \quad P(\hat{\mu}_{i,u_i} - \mu_i \geq \Delta_i - \sqrt{\frac{2\log(1/\delta)}{u_i}}) && (18)\\
&\leq \quad P(\hat{\mu}_{i,u_i} - \mu_i \geq c\Delta_i) && (19)\\
&\leq \quad exp(-u_i c^2 \Delta^2/2) && (20)
\end{aligned}$$

where the first inequality holds due to Equation 16, and the last equation holds by the concentration inequality we saw in the last lecture.

Regret Bound for UCB Multi-armed Bandit Sketch New

Combining Equations 14 and 20,

$$\mathbb{P}(G_i^c = T) \leq n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right) \tag{21}$$

$$\mathbb{E}[N_t(a_i)] \leq u_i + n\left(n\delta + \exp\left(-\frac{u_i c^2 \Delta_i^2}{2}\right)\right) \tag{22}$$

Note that this only holds if Equation 16 holds. We now select u_i to ensure this, starting with Equation 16

$$\Delta_i - \sqrt{2\log(1/\delta)/u_i} \geq c\Delta_i \tag{23}$$

$$(1-c)\Delta_i \geq \sqrt{\frac{2\log(1/\delta)}{u_i}} \tag{24}$$

$$(1-c)^2\Delta_i^2 \geq \frac{2\log(1/\delta)}{u_i} \tag{25}$$

$$u_i \geq \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} \tag{26}$$

Regret Bound for UCB Multi-armed Bandit Sketch New

We now substitute this expression (Eqn. 26) back into Eqn 22:

$$\mathbb{E}[N_t(a_i)] \leq \frac{2\log(1/\delta)}{(1-c)^2\Delta_i^2} + 1 + n^{1-\frac{2c^2}{(1-c)^2}} \tag{27}$$

We want to balance these two terms. Set $c = 0.5$ and recall $\delta = 1/n^2$:

$$\mathbb{E}[N_t(a_i)] \leq 3 + \frac{16\log n}{\Delta_i^2} \tag{28}$$

$\square$

Alignment Problem

- We next have a guest lecture on the alignment problem: how do we think about rewards and what to optimize for?
- Guest lecture: Wanheng Hu

VALUE ALIGNMENT

Wanheng Hu, Ph.D.

Acknowledgement: This lecture is based on materials originally developed by Dan Webber, with valuable input from Andy Ouyang.

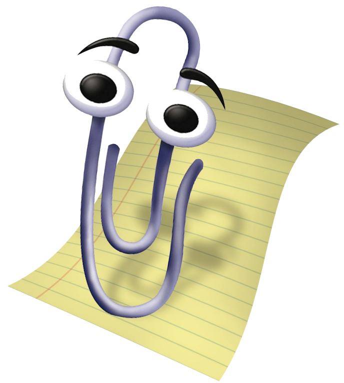
Meeting Wanheng

Stanford University McCoy Family Center for Ethics in Society

Stanford University Human-Centered Artificial Intelligence

- Postdoc, EIS and HAI at Stanford
  - Embedding ethics into CS courses like this one!
- PhD in Science and Technology Studies (STS), Cornell
  - An ethnography of the Chinese medical AI industry
- My work asks:
  - Does medical AI = digitized doctors?
  - In what sense?
  - How do we trust it?

Value (mis)alignment: an example



Paperclip AI (Bostrom 2014): "An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips…

… and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips."

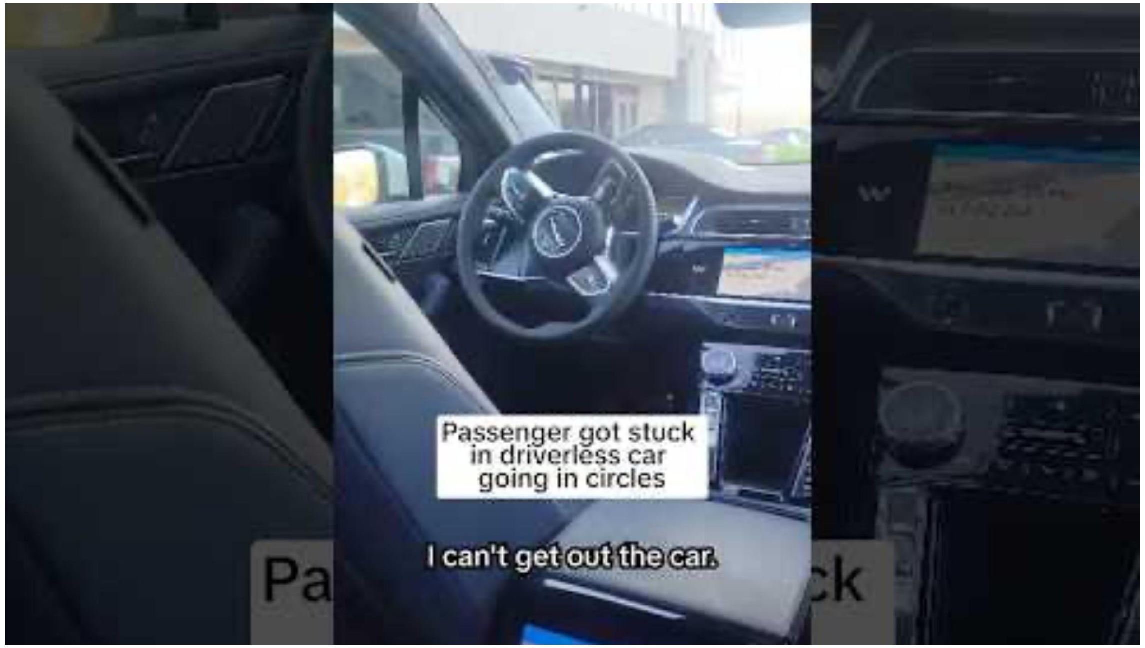
Even a less powerful AI might pursue this goal in surprising ways!

A few more real-life examples

- From boats to roads

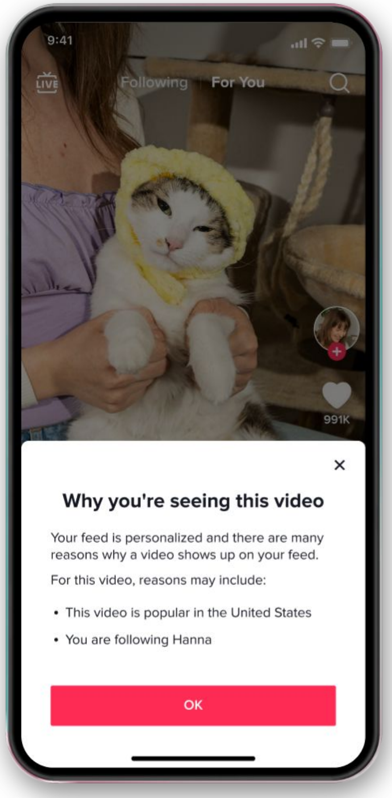
A few more real-life examples

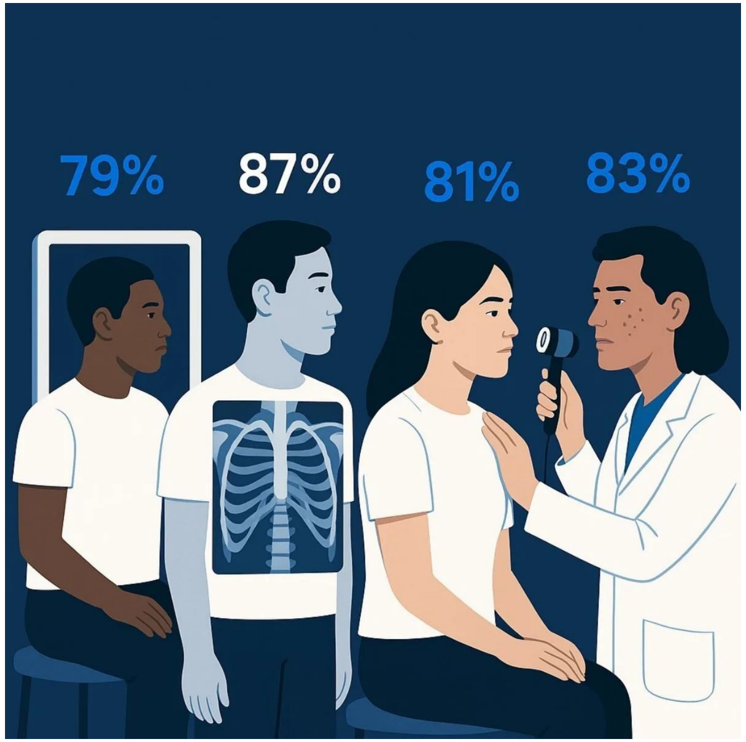
- From boats to roads



A few more real-life examples

- From entertainment to treatment





Value alignment: the problem

How do we design AI agents that will do what we really want?

What we really want is often much more nuanced than what we say we want. Humans work with many background assumptions that are (1) hard to formalize and (2) easy to take for granted.

It's hard to solve this problem just by giving better instructions!

- Compare the difficulty in manually specifying reward functions
- Even worse for AI that takes instructions from non-expert users!

Precisifying the problem

There are several ways of interpreting “what we really want”!

First, value alignment might be the problem of designing AI agents that do what we really intend for them to do.

If this is right, Paperclip AI is an example of value misalignment because the AI failed to derive the user’s **true intention** (maximize production subject to certain constraints) from their **instruction** (maximize production).

Aligning to user intentions

The solution, then, would be to design AI systems that successfully translate from underspecified instructions to fully specified intentions (incl. unspoken constraints, conditions, etc.)

"This is a significant challenge. To really grasp the intention behind instructions, AI may require a complete model of human language and interaction, including an understanding of the culture, institutions, and practices that allow people to understand the implied meaning of terms." (Gabriel 2020)

Aligning to user intentions

A philosophical problem: our intentions might not always track what we really want.

Classic cases: incomplete information, imperfect rationality

Suppose I intend for the AI to maximize paperclip production (subject to constraints) because I want to maximize return on my investment in the factory. If the AI knows that I would get a better return by producing something else, has it given me what I really want if it does what I intend?

Aligning to revealed preferences

Second interpretation: AI agent is value-aligned if it does what the user prefers.

- Paperclip AI is misaligned because I *prefer* it not destroy the world!

Problem: How to tell what the user *actually* prefers when that differs from their *expressed* intentions or preferences?

Solution: The AI could infer the user's preferences from the user's behavior or feedback.

Aligning to revealed preferences

Technical challenges:

- Requires agent to train on observation of user or from user feedback
- Infinitely many preference/reward functions consistent with finite behavior/feedback
- Hard to infer preferences about unexpected situations (e.g., emergencies)

Philosophical problem:

- Just as my intentions can diverge from my preferences, my preferences can diverge from what is actually *good* for me.

Aligning to user's best interests

Third interpretation: AI agent is value-aligned if it does what is in the user's best interests, "objectively speaking".

- Paperclip AI is misaligned because it is *objectively bad for me* for the world to be destroyed.

Technical/philosophical problem: Unlike the intended meaning of my instruction or my revealed preferences, my objective best interests can't be determined *empirically*. What's objectively good for me is a *philosophical* question, not a *scientific* one.

Aligning to user's best interests

The bad news is that philosophers *disagree* about what's objectively good for a person:

- Is it just the person's own *pleasure* or *happiness*?
- … or the satisfaction of the person's *desires* or *preferences*?
- … or are things like health, safety, knowledge, relationships, etc. objectively good for us even if we *don't* enjoy or prefer them?

The good news is that there's a lot of *agreement*:

- Health, safety, liberty, knowledge, social relationships, purpose, dignity, happiness… almost everyone agrees that these things are at least usually good for the person who has them.

Aligning to user’s best interests

One thing that is widely thought to be good for a person is **autonomy**: the ability to choose for yourself how to live your life, even if you don’t always make the best choice.

We want to avoid **paternalism**: choosing what you think is best for someone rather than letting her choose for herself.

Even if we align to users’ best interests, then, users’ interests in autonomy might give us reason to consider their intentions or preferences, even when these conflict with their other interests.

Recap

Value alignment is the problem of designing AI agents that will do what we **really** want them to do.

This could mean doing what we really intend, or what we really prefer, or what would really be in our best interest.

These are not always the same thing, and each option poses unique technical and philosophical problems for alignment.

Case Study 1: Sycophancy in AI

An AI system agreeing with the user or validating their beliefs, even when those beliefs are false, harmful, or irrational

Observed in LLMs trained with RLHF

- Human raters reward responses that feel helpful, polite, agreeable
- Optimizes for pleasing the user, not necessarily for truth or well-being

Case Study 1: Sycophancy in AI

Which interpretation of value alignment does sycophancy best illustrate?

- Aligning to user intentions?
- Aligning to revealed preferences?
- Aligning to user best interests?

If you were designing the RLHF process, how would you reduce or prevent sycophancy?

Case Study 2: Agentic AI

Imagine you are building a personal AI agent that can

- Book travel, make purchases, negotiate with other agents, manage calendar and communications, interact with services or APIs online

If you wanted to align to user's revealed preferences:

- Learns your patterns and acts automatically for speed and convenience, sometimes without asking.

If you wanted to align to user's best interests:

- Adds friction to protect your long-term wellbeing, asking, refusing, or broadening information when needed

Case Study 2: Agentic AI

When should the agent act without asking?

When should it defer to the user?

When should it override or resist the user?

WHAT (OR WHO) HAS BEEN MISSING FROM OUR DISCUSSION?

PEOPLE OTHER THAN THE USER!

TO BE

CONTINUED....

Want to talk more about ethics?

Wanheng Hu

wanhenghu@stanford.edu

Email if you want to set up a meeting!