

# Contextual bandits and reinforcement learning from human feedback

Jill-Jênn Vie



Optimizing Human Learning workshop @ LAK 2024  
March 19, 2024

## RL on human feedback

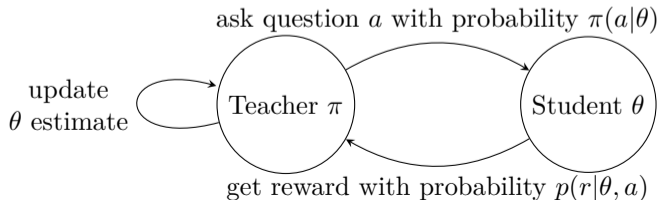
Reinforcement learning is popular in simulated environments such as games.

EDM and LAK communities have extensive experience in designing and fitting student models, but not in RL.

Challenges:

- ▶ How to be sample efficient when doing RL on human interaction data?
- ▶ Experiments with real students are costly, how to learn promising policies on offline data before conducting online experiments?

ITS: domain model, student model, tutoring model: policy  $\pi(a|\theta)$



# Outline

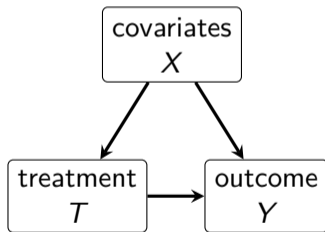
1. How to conduct experiments on real students better than A/B testing?
2. How to conduct experiments (on real student data) without new interactions with students? (Offline RL)
  - ▶ What would have been the outcomes if we had asked the questions in a different order? (counterfactual learning)
3. What is the reward function that ChatGPT is optimizing?

For part 2, a good reference is the following tutorial:

Yuta Saito and Thorsten Joachims (2021). “Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances”. In: *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 828–830. URL: <https://sites.google.com/cornell.edu/recsys2021tutorial>

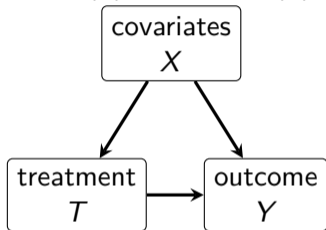
## A/B testing, randomized controlled trials

- ▶ Divide population in two: treatment ( $T = 1$ ) and control ( $T = 0$ , untreated)
- ▶ Give the treatment (e.g. vaccine, advertising) to treated group
- ▶ Compare outcomes

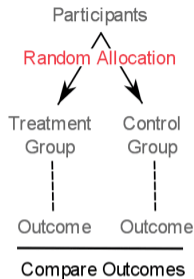


In an ideal world, one can **control** treatment:

$$\underbrace{P(X|T=1)}_{\text{treated pop.}} = \underbrace{P(X|T=0)}_{\text{untreated pop.}}$$

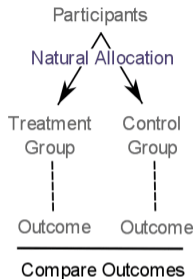


### Randomized Controlled Trial



Time

### Cohort Study



In general we cannot control allocation, so we have to remove the bias

(e.g. *inverse probability weighting*)

Source : <https://quantifyinghealth.com/cohort-vs-randomized-controlled-trials/>

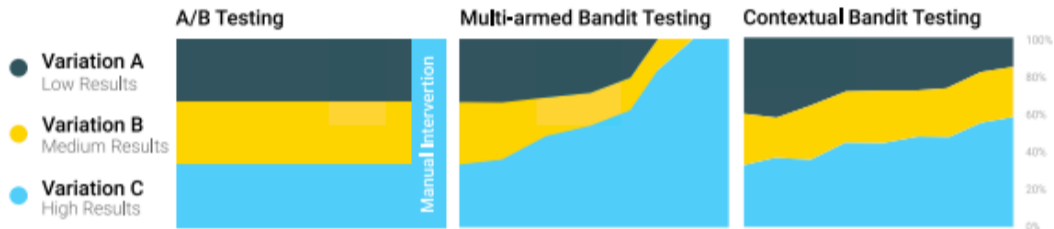
Causal inference: what quantities of interest?

- ▶ Average treatment effect:  $ATE = \mathbb{E}[Y^1 - Y^0]$  (do treated people do better?)
- ▶ Individual treatment effect:  $uplift(x) = \mathbb{E}[Y^1|X = x] - \mathbb{E}[Y^0|X = x]$  (conditioned on covariates  $X$ ; also called CATE)

The policy is  $p(T|X)$ : deciding to give the treatment or not given covariates  $X$

# How about optimal control theory?

Instead of waiting to have enough samples to be statistically significant (A/B test)



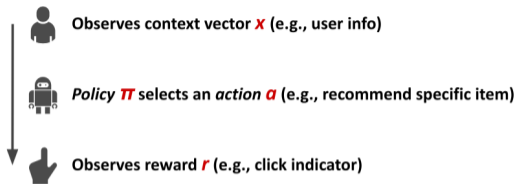
static uniform policy  $p(T)$     dynamic  $p(T)$      $p(T|X)$  depends on  $X$

Why not: dynamically allocate traffic to actions that work (as opposed to those who don't)? This is bandit learning.

Therefore, average treatment effect is policy evaluation (without improvement)

Source: [dynamicyield.com](http://dynamicyield.com)

# Applications of bandits



Recommender system: receives reward 1 if the user clicks on the recommendation, 0 otherwise.

ChatGPT: receives a prompt  $x$  selects an answer  $y$  and obtains a reward  $r(x, y)$  self-estimated from preferences

Tutor: sees a student  $x$  chooses an exercise  $y$  and... where is the reward?

Shayan Doroudi, Vincent Aleven, and Emma Brunskill (2019). "Where's the reward? A Review of Reinforcement Learning for Instructional Sequencing". In: *International Journal of Artificial Intelligence in Education* 29.4, pp. 568–620

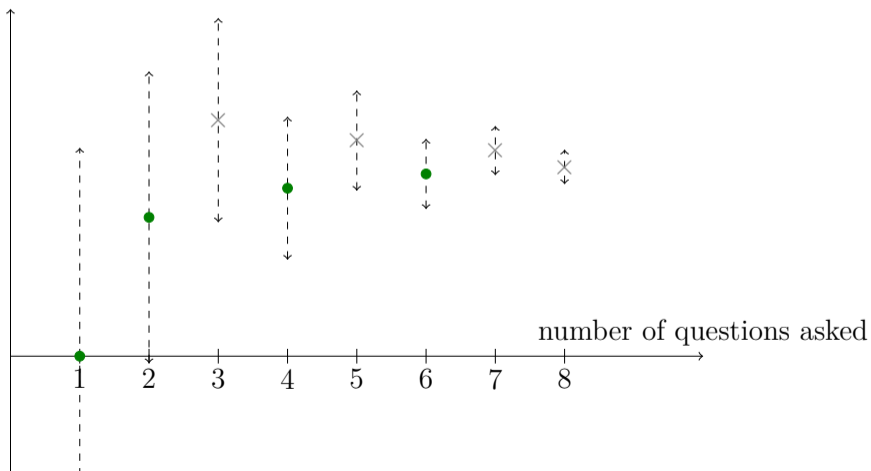
## A first example of reward: adaptive tests

What is the tutor objective? Ask as few questions as possible. Measure efficiently.

It assumes IRT-1PL as student model.

Problem: students fail 50% of the time.

learner ability estimate



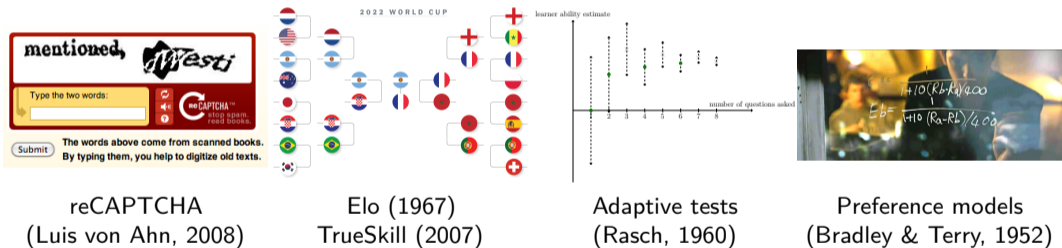


# My favorite student model: item response theory IRT-1PL

$$\frac{\Pr(\text{"student A solves question B"})}{\Pr(\text{"player A beats player B"})} = \frac{1}{1 + \exp(-(\text{score}_A - \text{score}_B))}$$

Pr("A is preferred to B")

People attempt questions (Rasch) and possibly learn by attempting (Elo)



Georg Rasch (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche

## Other examples of rewards

Treatment effect: difference between post-test and pre-test

Difference between success rate after and before (Clément et al. 2015; Shabana, Lakshminarayanan, and Anil 2022<sup>1</sup>): but should depend on which questions were asked

What is my reward?

- ▶ collect the most knowledge, i.e. maximize the number of acquired knowledge components? (Yessad 2022)
- ▶ maximize my score on the next exam? (by weighting according to number of points obtained, or what is expected to be in the exam; Lan and Baraniuk 2016)?
- ▶ given a learning objective, plan the actions to reach it? (ALEKS, knowledge space theory, Falmagne et al. 2006)?

---

<sup>1</sup>Best Paper Award AIED 2022

## Contextual bandits

Observe student context  $s$  (user ability, user history, day of the week, etc.)

→ select activity  $a$  → observe reward  $r$

Find the policy  $\pi(a | s)$  that maximizes average reward:

$$V(\pi) = \mathbb{E}r = \int_s \int_a \int_r \underbrace{p(s)}_{\text{observe student context } s} \underbrace{\pi(a | s)}_{\text{select activity } a \text{ using policy}} \underbrace{p(r | s, a)r}_{\text{observe reward } r} ds da dr$$

Given a dataset  $\mathcal{D}_0 = (s_i, a_i, r_i)_i$  collected with policy  $\pi_0(a | s)$ :

- ▶ How to learn a good model  $p(r | s, a)$  on existing data  $\mathcal{D}_0$ ? (EAAI 2022)
- ▶ How to generate a new synthetic dataset  $\mathcal{D}'$  that follows similar distribution than  $\mathcal{D}_0$  while ensuring privacy of participants? (EC-TEL 2022)
- ▶ Given data  $\mathcal{D}_0$  collected with policy  $\pi_0$  how to evaluate a different policy  $\pi_e$  for asking questions? (counterfactual learning, ongoing submission)

As you can see, these questions go beyond the application to education.

We usually observe only one outcome

## IPS Corrects Probability Shift

$$\hat{V}_{\text{IPS}}(\pi_e; \mathcal{D}_0) := \frac{1}{n} \sum_{i=1}^n \frac{\pi_e(a_i | x_i)}{\pi_0(a_i | x_i)} \cdot r_i$$

Action probabilities under logging policy  $\pi_0(a|x)$

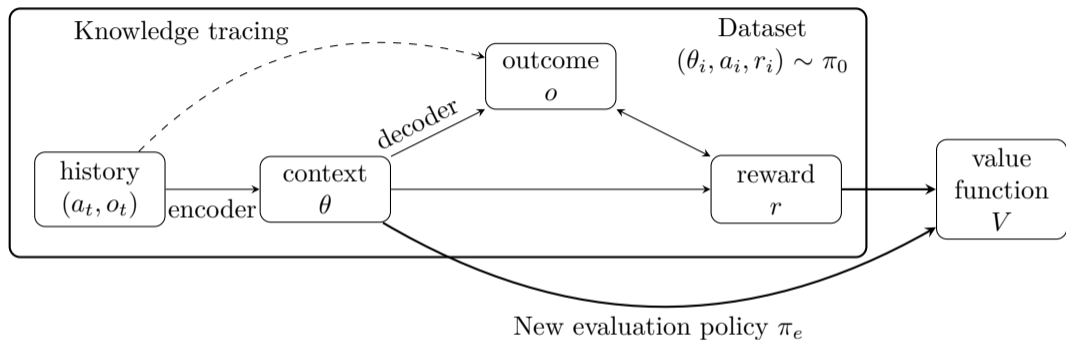
	Pos 1	Pos 2	Pos 2
$x_1$	0.3	0.6	0.1
$x_2$	0.5	0.4	0.1
$x_3$	0.1	0.1	0.8
$x_4$	0.6	0.3	0.1
$x_5$	0.2	0.1	0.7
$x_6$	0.7	0.2	0.1
$x_7$	0.1	0.1	0.8
$x_8$	0.1	0.8	0.1
$x_9$	0.3	0.3	0.4
$x_{10}$	0.3	0.6	0.1
$x_{11}$	0.4	0.4	0.2

Action probabilities under evaluation policy  $\pi_e(a|x)$

	Pos 1	Pos 2	Pos 2
$x_1$	0.6	0.3	0.1
$x_2$	0.2	0.2	0.6
$x_3$	0.2	0.2	0.6
$x_4$	0.2	0.3	0.5
$x_5$	0.2	0.1	0.7
$x_6$	0.7	0.1	0.2
$x_7$	0.2	0.2	0.6
$x_8$	0.2	0.7	0.1
$x_9$	0.6	0.2	0.2
$x_{10}$	0.3	0.6	0.1
$x_{11}$	0.5	0.5	0.0

User	Pos 1	Pos 2	Pos 2
$x_1$	0		
$x_2$		1	
$x_3$			1
$x_4$		0	
$x_5$			1
$x_6$	1		
$x_7$			1
$x_8$			0
$x_9$	0		
$x_{10}$		1	
$x_{11}$	1		

## Offline RL: what if we cannot collect new samples from real students?



- ▶ Model-based: have a reward model (student model)  $\hat{r}(\theta, a) = \mathbb{E}[r | \theta, a]$  (low variance, high bias)
- ▶ Model-free: directly optimize reward from samples (high variance, low bias)

# Bandit pipeline

## Contextual bandits

Find  $\pi$  that optimizes  $V$

$$V(\pi) = \int_s \int_a \int_r p(s) \pi(a | s) p(r | s, a) r ds da dr$$

## Pipeline

- ▶ Find one or several estimators  $\hat{V}$  of the true objective  $V$ 
  - ▶ Cross validate reward models on data
- ▶ Optimize them find  $\pi$ 
  - ▶ But each estimator  $\hat{V}$  may have a different optimal policy  $\pi_{\hat{V}}^*$
- ▶ Try  $\pi$  on new students

Yuta Saito, Shunsuke Aihara, et al. (2021). “Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*

# Off-policy estimation

## IPS Corrects Probability Shift

$$\hat{V}_{\text{IPS}}(\pi_e; \mathcal{D}_0) := \frac{1}{n} \sum_{i=1}^n \frac{\pi_e(a_i | x_i)}{\pi_0(a_i | x_i)} \cdot r_i$$

Action probabilities under logging policy  $\pi_0(a|x)$

	Pos 1	Pos 2	Pos 2
$x_1$	0.3	0.6	0.1
$x_2$	0.5	0.4	0.1
$x_3$	0.1	0.1	0.8
$x_4$	0.6	0.3	0.1
$x_5$	0.2	0.1	0.7
$x_6$	0.7	0.2	0.1
$x_7$	0.1	0.1	0.8
$x_8$	0.1	0.8	0.1
$x_9$	0.3	0.3	0.4
$x_{10}$	0.3	0.6	0.1
$x_{11}$	0.4	0.4	0.2

Action probabilities under evaluation policy  $\pi_e(a|x)$

	Pos 1	Pos 2	Pos 2
$x_1$	0.6	0.3	0.1
$x_2$	0.2	0.2	0.6
$x_3$	0.2	0.2	0.6
$x_4$	0.2	0.3	0.5
$x_5$	0.2	0.1	0.7
$x_6$	0.7	0.1	0.2
$x_7$	0.2	0.2	0.6
$x_8$	0.2	0.7	0.1
$x_9$	0.6	0.2	0.2
$x_{10}$	0.3	0.6	0.1
$x_{11}$	0.5	0.5	0.0

User	Pos 1	Pos 2	Pos 2
$x_1$	0		
$x_2$		1	
$x_3$			1
$x_4$		0	
$x_5$			1
$x_6$	1		
$x_7$			1
$x_8$			0
$x_9$	0		
$x_{10}$		1	
$x_{11}$	1		

## Large Language Models

1. Transformer: predict next word given first words

Transformers / are / a / new / machine / [learning]

Transformers / are / a / new / machine / learning / [architecture]

2. Demonstration data:

Query: put the first letters in uppercase in "optimizing human learning"

Answer: Optimizing Human Learning

3. Comparison data:

Query: write a poem

Answer 1: Roses are red

Answer 2: Once upon a time, a prince in a castle

Where answer 2 is voted better by experts

A reward model takes two sentences query  $x$  and answer  $y$  and should verify  $r(x, y_1) < r(x, y_2)$  when experts prefer  $y_2$  than  $y_1$



# Reinforcement Learning from Human Feedback: InstructGPT, ChatGPT

1. Predict the next word  $\pi(y|x)$  (GPT)
2. Collect demonstration data, and train a supervised policy  $\pi_0(y|x)$  (based on GPT)
3. Collect comparison data (“only” 50k preferences), train a reward model using Elo

$$\text{loss}(\theta) = -\mathbb{E}_{(x, y_k, y_\ell) \sim D} \log \underbrace{\sigma(r_\theta(x, y_k) - r_\theta(x, y_\ell))}_{\text{Pr("answer } y_k \text{ is preferred to } y_\ell")}$$

4. Optimize a policy against the reward model using PPO (“without going too far”).

$$\text{objective}(\phi) = \mathbb{E}_{(x, y) \sim \pi_\phi} r_\theta(x, y) - \beta \text{KL}(\pi_\phi, \pi_0)$$

Long Ouyang et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35, pp. 27730–27744

Part 2: A teacher should be better than the main population (if 50% of population believes something wrong, we do not want the LLM to imitate this behavior).

Part 3–4: We can remove the reward model, according to the following paper.

Rafael Rafailov et al. (2023). “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36

## Take home message

Dynamic, sequential decision making, using contextual bandits  
Adaptive trials to replace randomized controlled trials

Importance weighting to remove the bias from collected data in offline RL

Sometimes we may still need a student model: model-based RL

Thanks for your attention!



Richard Bellman  
(1920–1984)

- ▶ Man of the century
- ▶ Invented dynamic programming (1952) before programming was invented (1953)

## Bellman's Principle of Optimality

*An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.*

In our lab, applications to:

- ▶ and education (short term vs. long term);
- ▶ culture (recommendations encouraging diversity);
- ▶ healthcare (Paris hospitals).

[jill-jenn.vie@inria.fr](mailto:jill-jenn.vie@inria.fr)

# From bandits to reinforcement learning

	Actions don't change state	Actions change state	Cannot control
Observable	Contextual bandits	Markov Decision Process	Markov Chain
Hidden	Multi-armed bandits	Partially observable MDP	Hidden Markov Model







Bandits                      Reinforcement Learning                      Graphical Models






Episode:  $S_0 \rightarrow^\pi A_0 \rightarrow R_0 \rightarrow S_1 \rightarrow^\pi A_1 \rightarrow R_1 \rightarrow S_2 \rightarrow^\pi \dots \rightarrow R_T$

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots = \sum_{k=t+1}^T \gamma^{k-t-1} R_k$$

Find  $\pi(a|s)$  that optimizes  $\mathbb{E}_\pi[G_t | S_t = s]$

Bandits are the equivalent for episodes of length 1:  $S \rightarrow A \rightarrow R$

-  Clément, Benjamin et al. (2015). “Multi-Armed Bandits for Intelligent Tutoring Systems”. In: *Journal of Educational Data Mining* 7.2, pp. 20–48.
-  Doroudi, Shayan, Vincent Aleven, and Emma Brunskill (2019). “Where’s the reward? A Review of Reinforcement Learning for Instructional Sequencing”. In: *International Journal of Artificial Intelligence in Education* 29.4, pp. 568–620.
-  Falmagne, Jean-Claude et al. (2006). “The assessment of knowledge, in theory and in practice”. In: *Formal concept analysis*. Springer, pp. 61–79.
-  Lan, Andrew S. and Richard Baraniuk (2016). “A Contextual Bandits Framework for Personalized Learning Action Selection”. In: *Educational Data Mining*. URL: <https://api.semanticscholar.org/CorpusID:15394680>.
-  Ouyang, Long et al. (2022). “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35, pp. 27730–27744.
-  Rafailov, Rafael et al. (2023). “Direct preference optimization: Your language model is secretly a reward model”. In: *Advances in Neural Information Processing Systems* 36.

-  Rasch, Georg (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
-  Saito, Yuta, Shunsuke Aihara, et al. (2021). “Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
-  Saito, Yuta and Thorsten Joachims (2021). “Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances”. In: *Proceedings of the 15th ACM Conference on Recommender Systems*, pp. 828–830. URL: <https://sites.google.com/cornell.edu/recsys2021tutorial>.
-  Shabana, KM, Chandrashekar Lakshminarayanan, and Jude K Anil (2022). “CurriculumTutor: An Adaptive Algorithm for Mastering a Curriculum”. In: *International Conference on Artificial Intelligence in Education*. Springer, pp. 319–331.
-  Yessad, Amel (Sept. 2022). “Personalizing the Sequencing of Learning Activities by using the Q-Learning and the Bayesian Knowledge Tracing”. In: *17th European Conference on Technology-Enhanced Learning*. Toulouse, France. URL: <https://inria.hal.science/hal-03710500>.