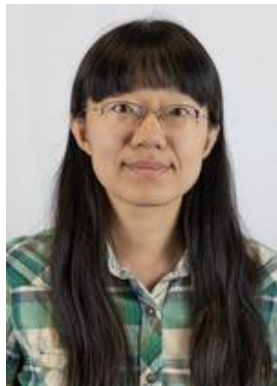





A reinforcement learning approach to personalized learning recommendation systems



Xueying Tang¹, Yunxiao Chen^{2*} , Xiaoou Li³, Jingchen Liu¹ and
Zhiliang Ying¹

¹Department of Statistics, Columbia University, New York, New York, USA

²Department of Psychology, Institute for Quantitative Theory and Methods,
Emory University, Atlanta, Georgia, USA

³School of Statistics, University of Minnesota, Minneapolis, Minnesota, USA

Reporter: Yingshi Huang

Introduction

- Personalized/adaptive learning



Feeling boredom because you already mastered the classroom material?



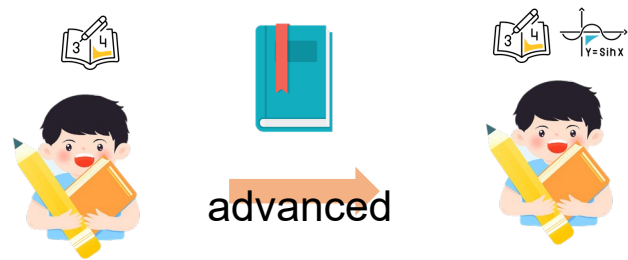
Experiencing stress because the teacher was teaching too fast for you?

Introduction

- Personalized/adaptive learning



already master
basic skill



- How to determine the tailored learning path for each learner?
 - Goal:
maximize the overall reward along the whole learning process for each learner
 - Key question:
makes decisions on what to learn at the next step

- How to determine the tailored learning path for each learner?
 - Three components:
 - Measurement model
(students' current knowledge profile)
 - Learning model
(the learning process: relationship between learning materials and changes of knowledge profiles)
 - Recommendation strategy
(the selection of learning materials)

- How to determine the tailored learning path for each learner?

- **Three components:**

Measurement model

(students' current knowledge profile)

Learning model (prior): complex & require large sample size to calibrate

(the learning process: relationship between learning materials and changes of knowledge profiles)

Recommendation strategy

(the selection of learning materials)



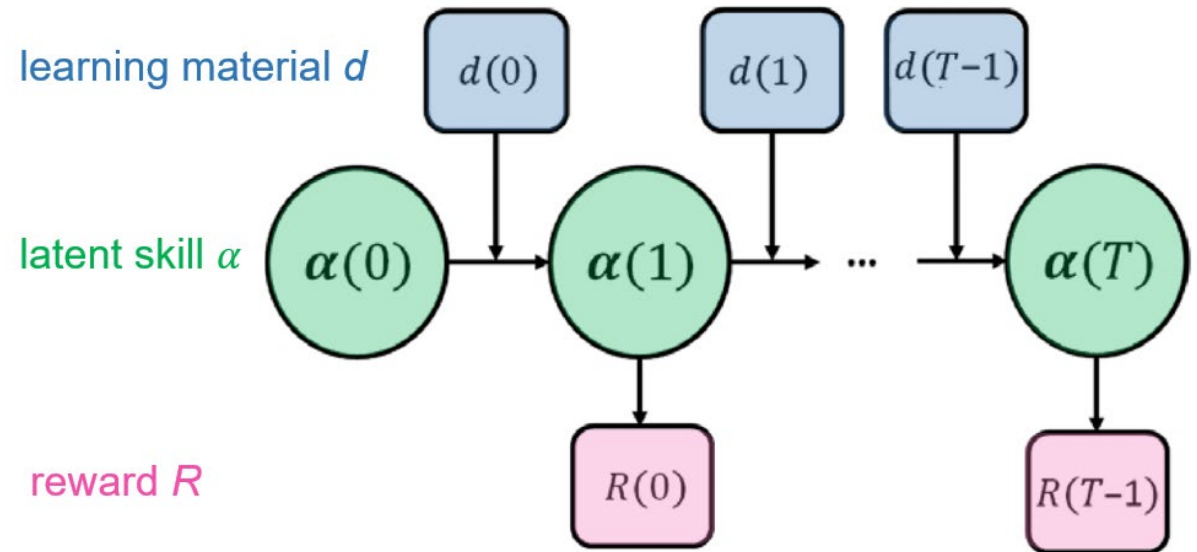
Purpose:

simultaneously build the learning model and recommendation strategy

Background

7

- K skills: $\alpha_1, \alpha_2, \dots, \alpha_k$ (mastery = 1, non-mastery = 0)
- T time epochs: $0, 1, \dots, T - 1$
- Learning material pool: \mathcal{D}
- Reward
 - the number of skills being mastered at learning stage t :
 - $R(t) = \sum_{k=1}^K [\alpha_k(t+1) - \alpha_k(t)]$
- the entire learning process:
 - $E(\sum_{t=0}^{T-1} R(t))$

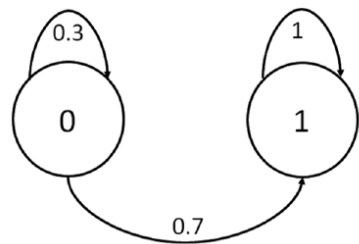


- Measurement model:
 - the probability of a specific response on item j : $P(Y_j = y|\alpha)$
 - diagnostic models (discrete) or multidimensional IRT (continuous)

- Measurement model:
 - the probability of a specific response on item j : $P(Y_j = y|\alpha)$
 - diagnostic models (discrete) or multidimensional IRT (continuous)

- Learning model:

- the effectiveness of each learning material
- a Markov chain (with no retrogression assumption): $P_d(\alpha(t+1) = \alpha | \alpha(t) = \tilde{\alpha})$



→ no arrow pointing from 1 to 0

$$P(\alpha(t+1) = 0 | \alpha(t) = 0) = 0.3$$

only depends on time t

- contain a large number of parameters: $|\mathcal{D}| \times 2^K$
(the number of learning materials \times all possible states of knowledge profiles)

- Recommendation strategy:
 - the probability that material d will be recommended at time t : policy π
 - $\pi_t(d) \geq 0$ & $\sum_{d \in \mathcal{D}} \pi_t(d) = 1$
- lower benchmark: $1/|\mathcal{D}|$
- upper benchmark: oracle strategy π^*
 - when no measurement error and learning model is known, that is, outperform any policy under imprecise information



approximate π^* by collecting students' learning data in a strategic way

Reinforcement Q-learning

11

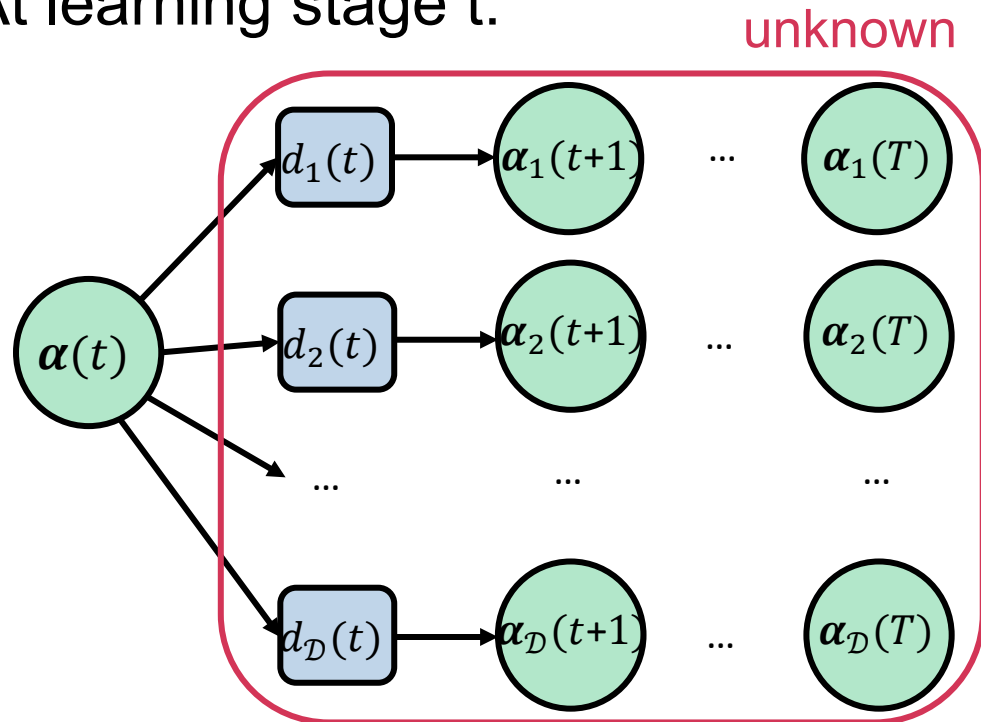
- Objective:
 - (1) bypass the estimation of learning model & (2) approximate π^*
 - how to optimize the policy without the learning model

Reinforcement Q-learning

12

- Objective:
 - (1) bypass the estimation of learning model & (2) approximate π^*
→ how to optimize the policy without the learning model

- At learning stage t :



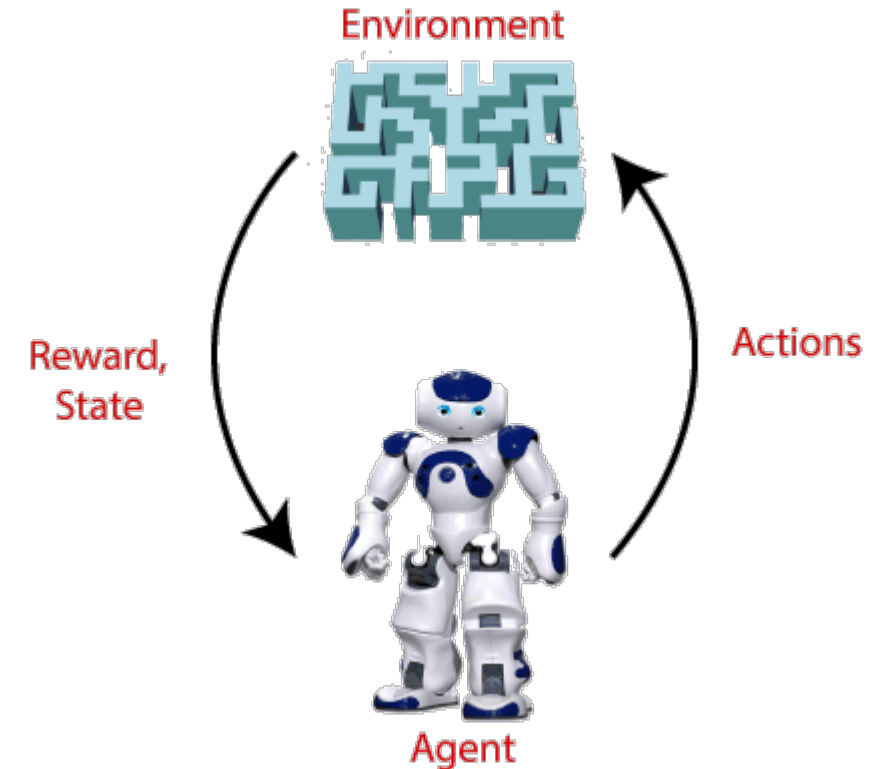
What would happen after selecting different learning materials is unknown

But we need to maximize the overall reward

Reinforcement Q-learning

13

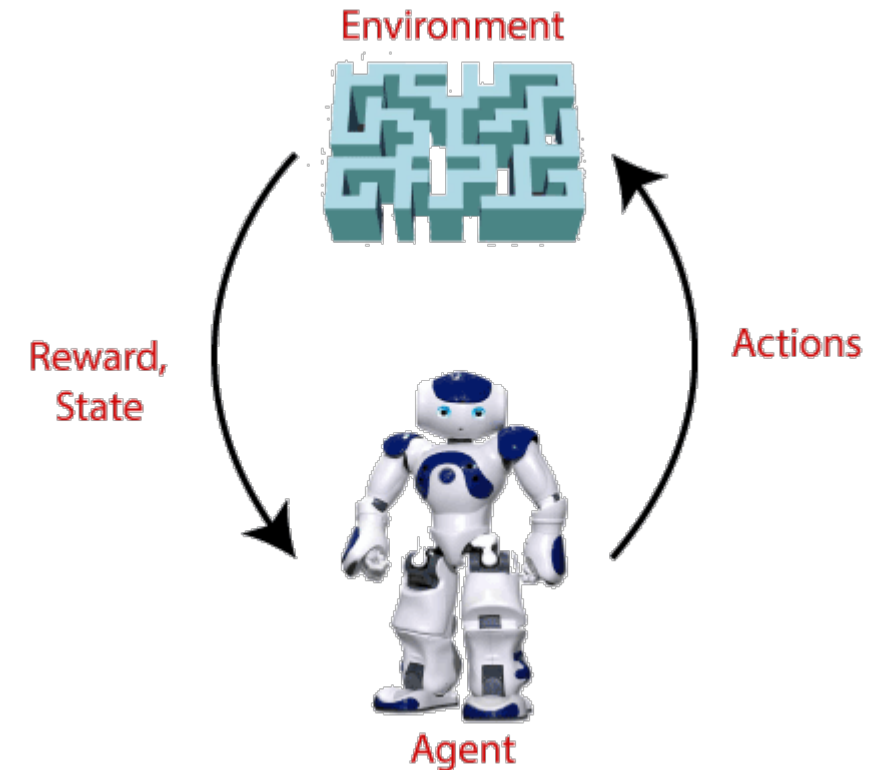
- The principle of reinforcement learning:
 - learn in an interactive environment by trial and error
 - find a suitable action model (sequential actions) that would maximize the total cumulative reward (a long-term goal) of the agent



Reinforcement Q-learning

14

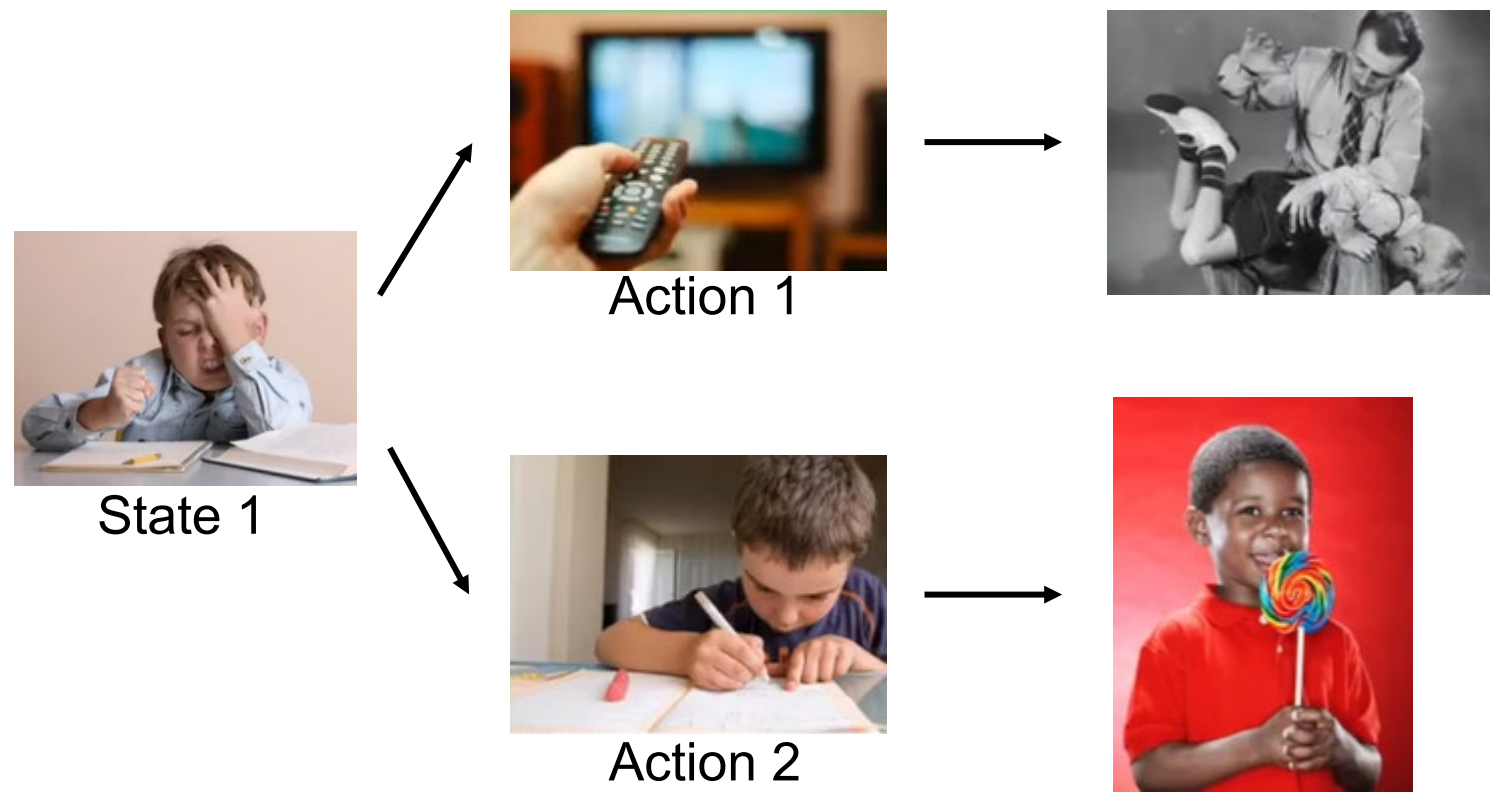
- The principle of reinforcement learning:
 - learn in an interactive environment by trial and error
 - find a suitable action model (sequential actions) that would maximize the total cumulative reward (a long-term goal) of the agent
- In this case:
 - Agent → learner
 - State → knowledge profile $\alpha(t)$
 - Action → selection of learning material
 - Environment → learning process
 - Reward → the changes of knowledge profile



Reinforcement Q-learning

- Q-learning algorithm:
 - Determine action sequence with Q table

	Action 1	Action 2
State 1	-5	10

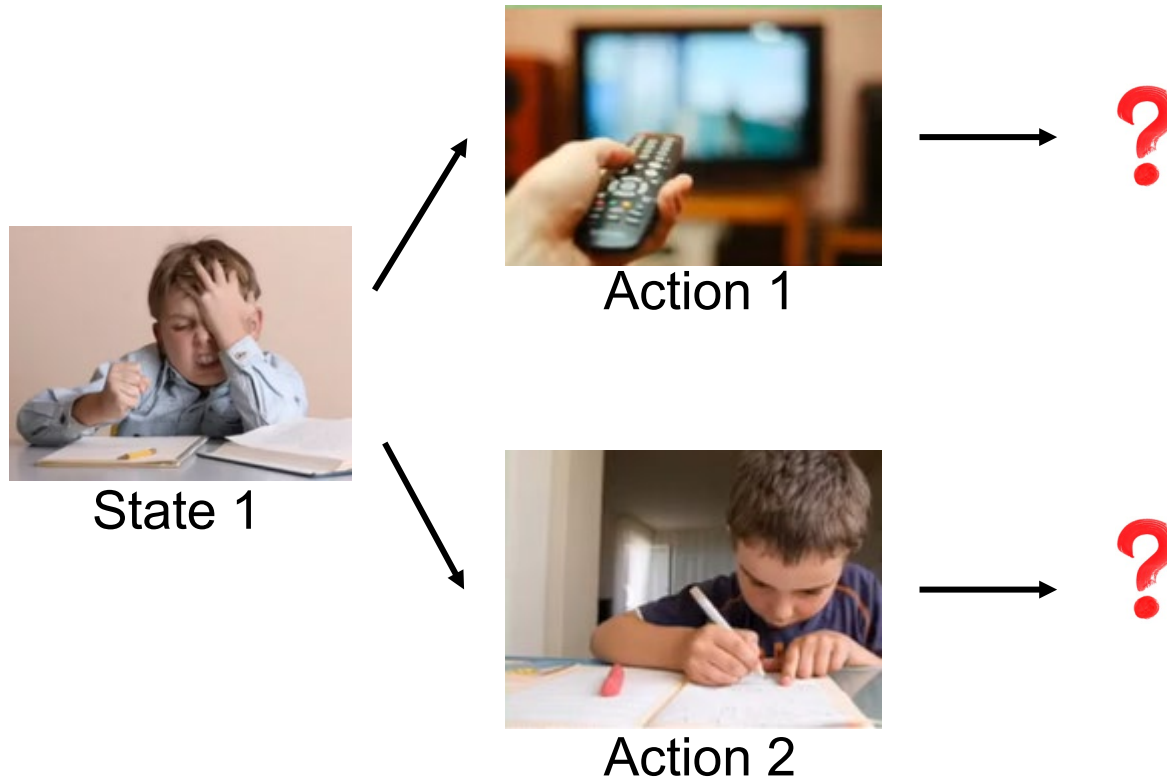


Reinforcement Q-learning

16

- Q-learning algorithm:
 - Determine action sequence with Q table

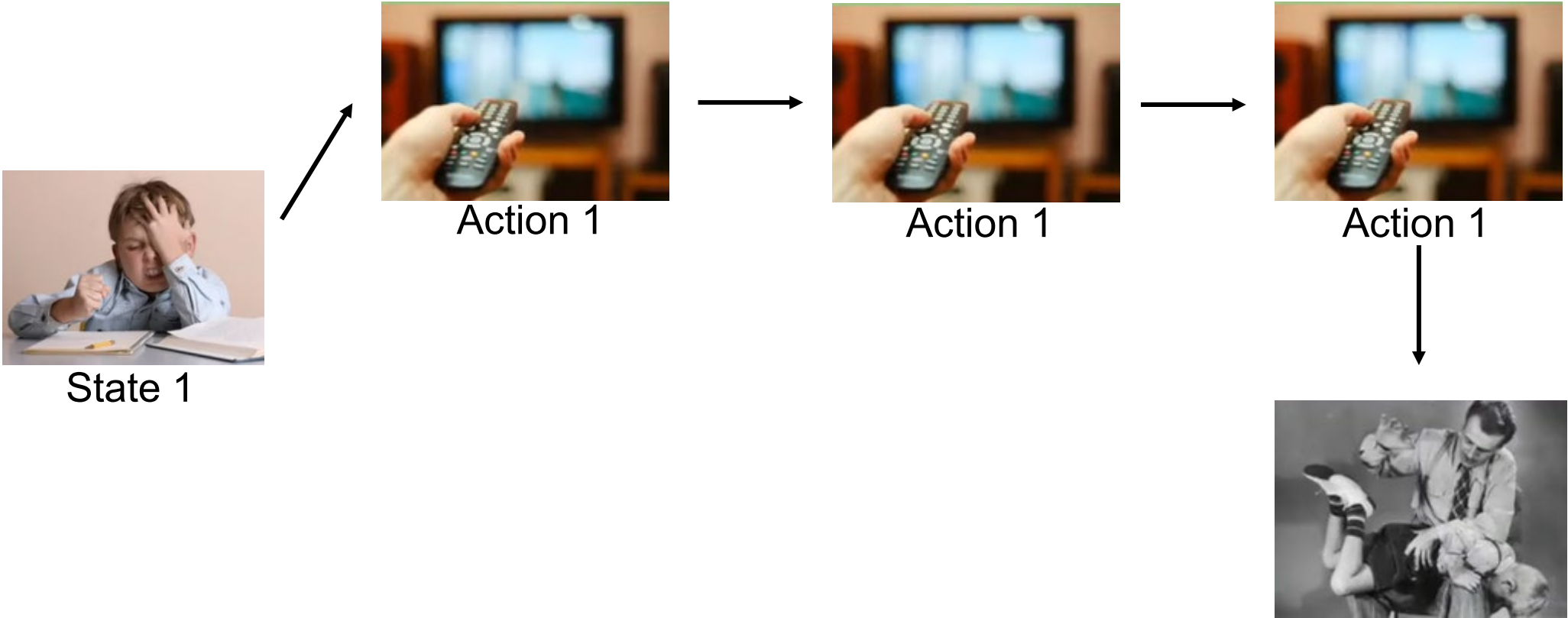
	Action 1	Action 2
State 1	0	0



Reinforcement Q-learning

- Q-learning algorithm:
 - Determine action sequence with Q table

	Action 1	Action 2
State 1	-5	0



- Q-learning algorithm:
 - Determine action sequence with Q table
 - The sum of the expected reward gained in the remaining training epochs

$$Q_t^*(\alpha, d) = E \left(\sum_{s=t}^{T-1} R(s) | \alpha, d, \pi^* \right)$$

- Maximize $Q_t^*(\alpha, d)$ to select the next learning material

$$d^* = \operatorname{argmax}_d Q_t^*(\alpha, d)$$

- Q-learning algorithm:
 - Determine action sequence with Q table
 - The sum of the expected reward gained in the remaining training epochs

$$Q_t^*(\alpha, d) = E \left(\sum_{s=t}^{T-1} R(s) | \alpha, d, \pi^* \right)$$

- Maximize $Q_t^*(\alpha, d)$ to select the next learning material
$$d^* = \operatorname{argmax}_d Q_t^*(\alpha, d)$$
- Example: two skills ($K = 2$) with two set of learning materials in two time epochs ($T = 2$)

α	$t = 0$		$t = 1$	
	$d = 1$	$d = 2$	$d = 1$	$d = 2$
(0, 0)	1.26	0.60	0.60	0.00
(1, 0)	0.70	0.91	0.00	0.70

Reinforcement Q-learning

20

- Q-learning algorithm:
 - Determine action sequence with Q table
 - The sum of the expected reward gained in the remaining training epochs

$$Q_t^*(\alpha, d) = E \left(\sum_{s=t}^{T-1} R(s) | \alpha, d, \pi^* \right)$$

- Maximize $Q_t^*(\alpha, d)$ to select the next learning material
$$d^* = \operatorname{argmax}_d Q_t^*(\alpha, d)$$
- Example: two skills ($K = 2$) with two set of learning materials in two time epochs ($T = 2$)

α	$t = 0$		$t = 1$	
	$d = 1$	$d = 2$	$d = 1$	$d = 2$
(0, 0)	1.26	0.60	0.60	0.00
(1, 0)	0.70	0.91	0.00	0.70



$T \times 2^K \times |\mathcal{D}|$
So complex!

Reinforcement Q-learning

21

- Q-learning algorithm:
 - $Q_t^*(\alpha, d)$ is approximated by a linear model

$$Q_t(\hat{\alpha}, d, \beta) = \sum_{l=1}^p \beta_l^{(td)} f_l(\hat{\alpha})$$

finite dimensional vector
functions summarizing features of $\hat{\alpha}$

→ from $T \times 2^K \times |\mathcal{D}|$ to $T \times p \times |\mathcal{D}|$ ($p \ll 2^K$)

- Q-learning algorithm:

- $Q_t^*(\alpha, d)$ is approximated by a linear model

$$Q_t(\hat{\alpha}, d, \beta) = \sum_{l=1}^p \beta_l^{(td)} f_l(\hat{\alpha})$$

finite dimensional vector
functions summarizing features of $\hat{\alpha}$

→ from $T \times 2^K \times |\mathcal{D}|$ to $T \times p \times |\mathcal{D}|$ ($p \ll 2^K$)

- Example: a main effect linear model

$$Q_t(\hat{\alpha}, d, \beta) = \beta_0^{(td)} + \sum_{k=1}^K \beta_k^{(td)} f_k(\hat{\alpha})$$

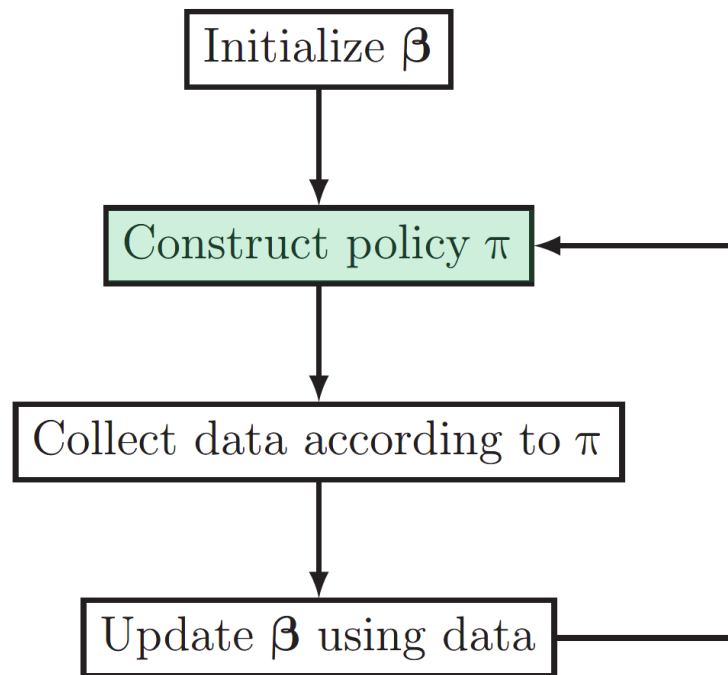
the posterior probability of $\alpha_k(t) = 1$

➡ Our problem becomes the estimation of β

Reinforcement Q-learning

23

- The estimation of β :
 - the balance between **exploration** (exploring new path) and **exploitation** (following the current “best” path)



$$\pi_t(d|\hat{\alpha}) = \frac{\exp(\gamma_1 Q_t(\hat{\alpha}, d', \beta))}{\sum_{d' \in \mathcal{D}} \exp(\gamma_1 Q_t(\hat{\alpha}, d', \beta))}$$

exploration parameter ≥ 0

$\gamma_1 = 0$: purely random

$\gamma_1 = \infty$: completely follows the current Q-function

$$\beta_l^{new} = \beta_l^{old} + \text{learning rate} \times \Delta_l$$

- Adaptive Learning aims to provide **tailored learning trajectory** for every individual
- **Three key components** in personalized learning
 - Measurement model, learning model, and recommendation strategy
- Facilitating a solution with **reinforcement Q-learning**
 - Determine an optimal action sequence that maximizes the long-term reward through collecting feedbacks from the environment

The End. Thanks
for **Listening!**



School of Education
& Information Studies

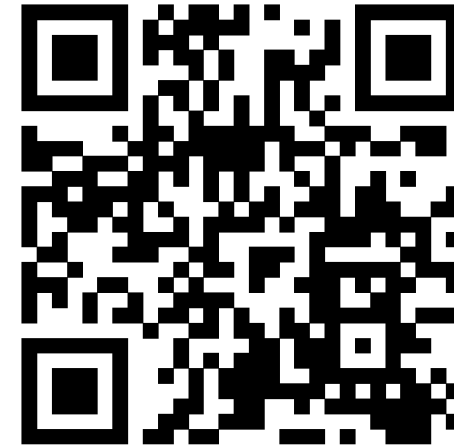
Thanks

감사합니다

Grazie

谢谢大家

Reporter: Yingshi Huang



Scan here to subscribe :)