CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 18: Agents That Learn

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

16TH MAR, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 18

- Introduction Segment
  - Information about Columbia, SC's open data
  - Recap/ Discussion of Lecture 17

- Main Segment
  - Reinforcement learning
  - Bayesian Optimization

- Concluding Segment
  - About Next Lecture – Lecture 19
  - Ask me anything

Introduction Segment

Columbia, SC's Open Data

Acknowledgements: Sung Jun Kim, Sylvia White

- Available Data
  - https://gis.columbiasc.gov/
  - Focus is on geo-mapping
  - Police records based on RMS standard (https://www.police1.com/police-products/police-technology/software/rms/). RMS data mainly contains public police records such as traffic stops, arrests and so on.
  - This is not very useful for understanding and solving civic issues like water, traffic, environment, food.
- Hackathon
  - https://www.onecolumbiasc.com/event/hack-for-sc/

City's Maturity on Open Data Maturity Scale

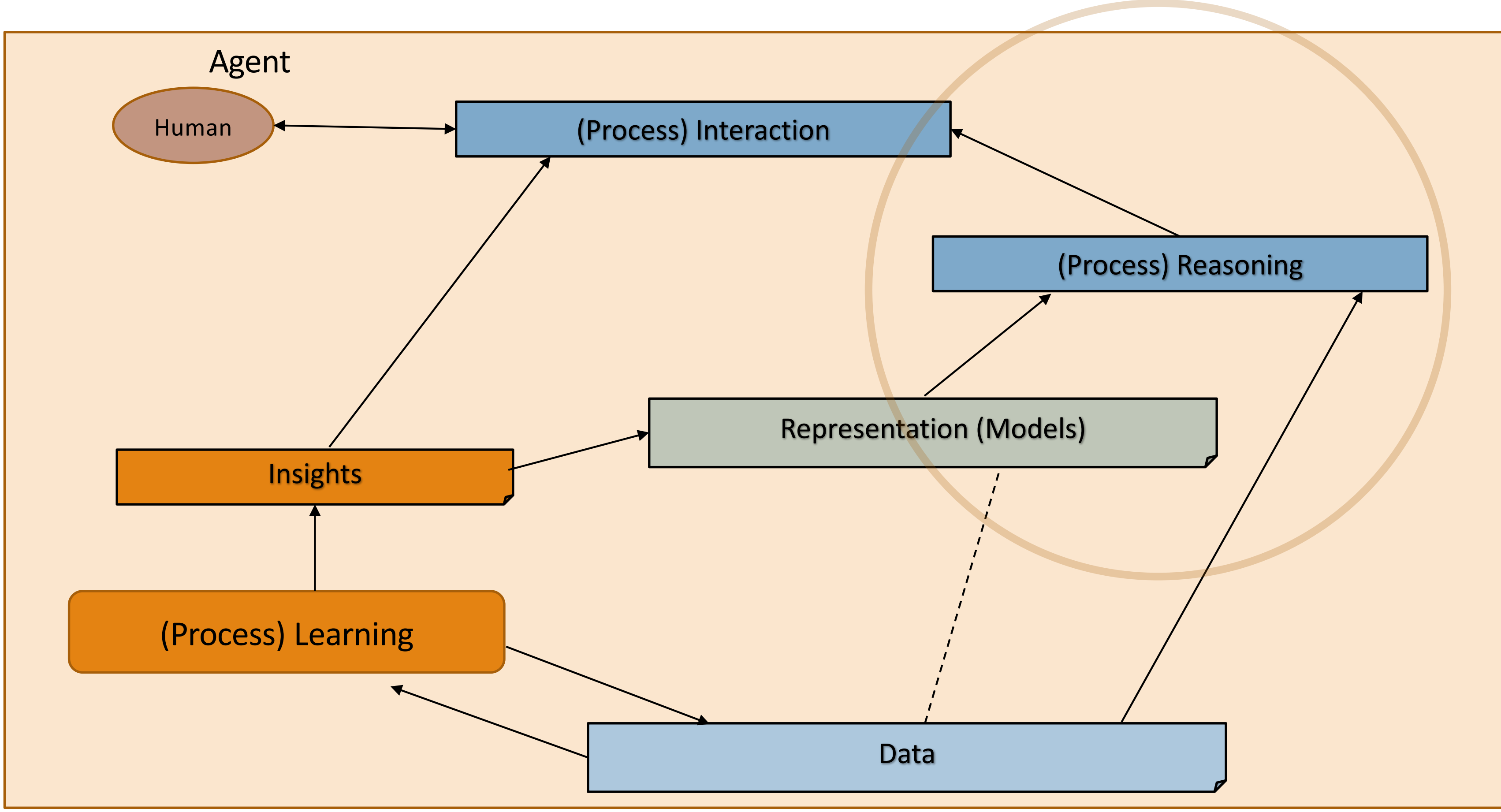
- 1- is no open data,
- 2- is some but not usable beyond a narrow focus (like geography),
- 3- where data is available across domains of governance
- 4- where people are engaged in using it for economic development widely

Columbia, SC is at level 2

Recap of Lecture 17

- Kind of uncertainties
- What is the best decision possible: Maximize Expectation
- Some methods
  - Bayesian methods
  - Utility theory
  - Markov Decision Processes

Relationship Between AI Processes



Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: **classification, clustering**, dimensionality reduction, anomaly detection, *neural methods*
- Predictive analysis
  - Predict about a new situation
  - **Methods**: **time-series**, *neural networks*
- **Prescriptive analysis**
  - What an agent should do
  - **Methods**: simulation, *reinforcement learning, Bayesian optimization*, **reasoning**

- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

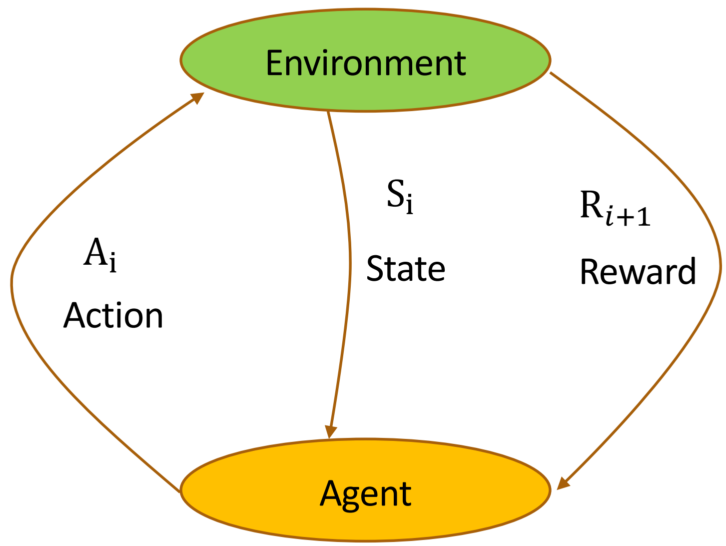
Example Situation – Course Selection

- A person wants to pass an academic program in two majors: A and B
- There are three subjects: A, B and C, each with three levels (*1, *2, *3). There are thus 9 courses: A1, A2, A3, B1, B2, B3, C1, C2, C3
- To graduate, at least one course at beginner (*1) level is needed in major(s) of choice(s), and two courses at intermediate levels (*2) are needed
- **Learning Agent**: *The student learns from their performance in earlier (e.g., level-1) courses, from others who have take courses or graduated, switches courses mid-semester*
- Answer questions
  - Q1: Should I switch my course in the middle of the program ?
  - Q2: Should I major in all the courses that the program has?
  - Q3: Should I drop dual major and focus on one? Which one?
  - …

Main Segment

Reinforcement Learning Setting

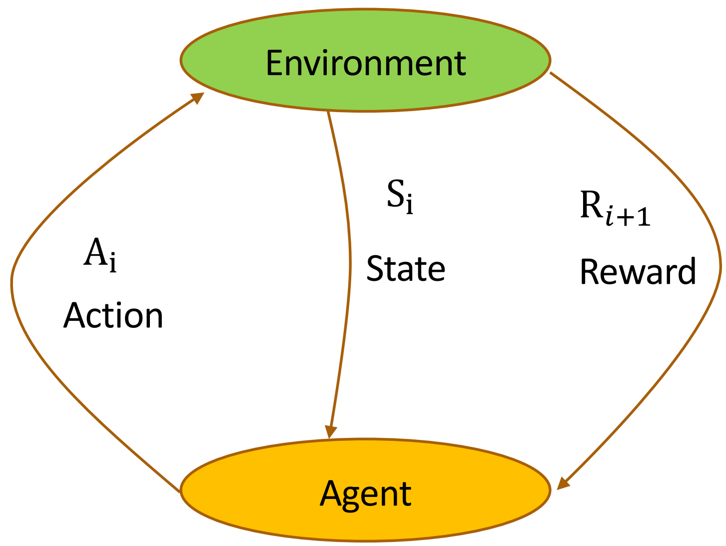
- An agent in an environment
- Agent
  - Can see state
  - Can take action
  - Will get rewards
- Precisely, at each time step i
  - In state S_i, agent takes action A_i
  - Based on state s_i and action a_i, the environment transitions to state S_{i+1} and outputs reward R_{i+1}
- **Objective**: learn mapping of states to actions so that the agent maximizes the reward from the environment.



Reinforcement Learning

- **Objective**: learn mapping of states to actions so that the agent maximizes the reward from the environment.

- **Output**
  - Deterministic: $a = \pi(s)$
  - Stochastic: $\pi(a|s) = P(A_i = a|S_i = s)$



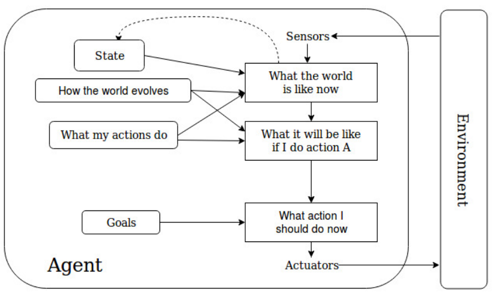
Comparison With Other Learning

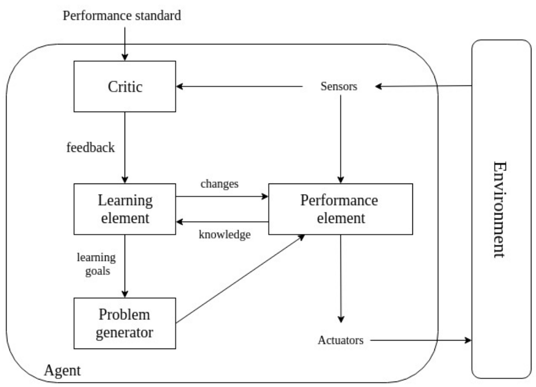
- Supervised learning
  - Training information: labels
  - Objective: learn (input-label) mapping
  - Goodness criteria: Reduce error = (Predicted label – Actual label)
- Reinforcement learning
  - Training information: reward functions
  - Objective: learn policy
  - Goodness criteria: maximal reward
- These two forms of learning are orthogonal – for different tasks

Image Source: Russell & Norvig, AI: A Modern Approach

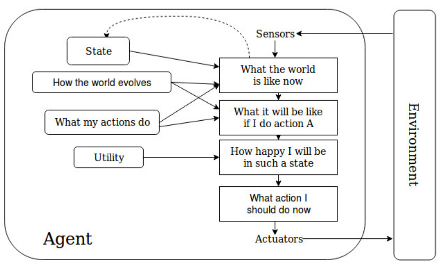
RL as a Learning-Based Agent

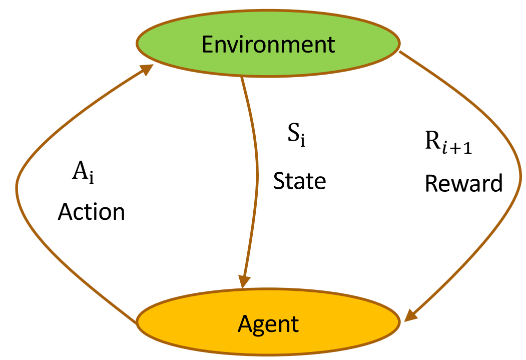
A general, alternative way of solving goal-based problems from just execution traces





Goal- and Utility-based Intelligent Agent





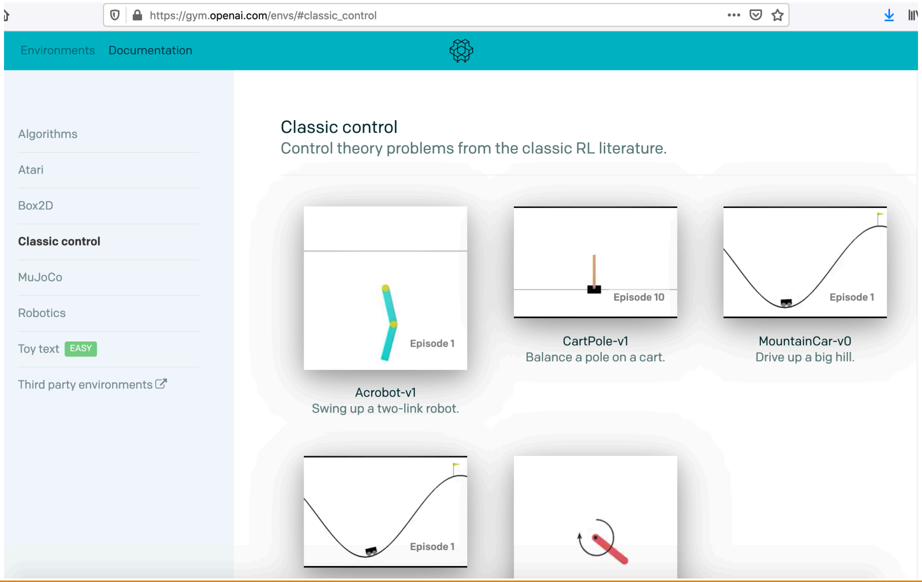
Model → Planning → Policy

(Generalization of plan)

Execution trace / Experience → RL → Policy

Exercise and Code – Gym RL

- RL using Open AI's Gym
  - https://gym.openai.com/
  - Environments: https://gym.openai.com/envs/#classic_control
- Exercise (5 mins):
  - Look at the various categories
  - Explore the videos



Exercise and Code – Gym RL

- RL using Open AI's Gym
  - https://gym.openai.com/

- Code: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l18-learning-agent/RL%20using%20Gym.ipynb

Source: Russell & Norvig, AI: A Modern Approach

Diversity in RL Problems

- Environment - accessible or inaccessible
  - Accessible: states can be identified with percepts
  - Inaccessible environment: agent has to learn and maintain representation of state to track environment
- Knowledge of effects of action and utility, or learn
- Rewards
  - Available for all states or only terminal states
  - Actual utility or hints of increase/ decrease
- Ability to execute actions - Active learner or passive learner
  - A passive learner simply watches the world going by, and tries to learn the utility of being in various states
  - An active learner can actions to explore unknown environment

Source: Russell & Norvig, AI - A Modern Approach

Passive RL

- **Input**
  - policy: π_i
  - // Has no knowledge Reward R(s) and Transition function P(s' |s, a)
- **Output**
  - Expected utility for each state, U(s)
- **Procedure**:
  - Execute a sequence of runs
  - At any instant, the agent knows only its current state and current reward, and the action it must take next. This action may lead it to more than one state, with different probabilities.
- **Expected Utility**

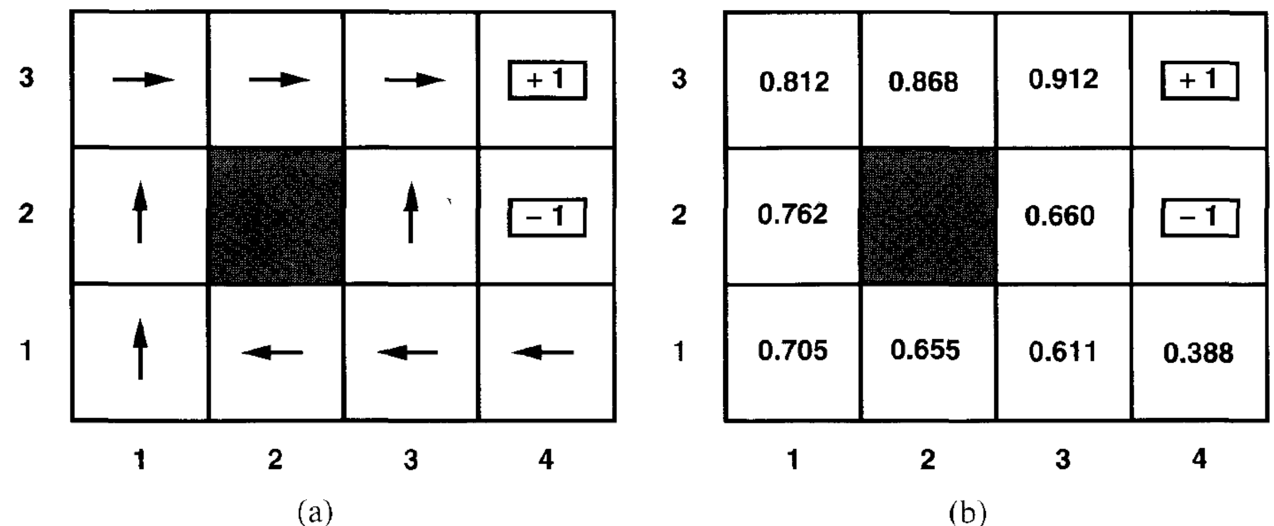
$$U^{\pi}(s) = E(\sum_{t\,=\,0}^{\text{inf}} \gamma^t R^t(s\,'\,))$$

Illustration

```
# Action Directions
north = (0, 1)
south = (0,-1)
west = (-1, 0)
east = (1, 0)

policy = {
    (0, 2): east,  (1, 2): east,  (2, 2): east,   (3, 2): None,
    (0, 1): north,                (2, 1): north,  (3, 1): None,
    (0, 0): north, (1, 0): west,  (2, 0): west,   (3, 0): west,
}
```

Policy: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l15-l16-l17-l18-agents/reinforcement_learning.ipynb



Input Policy and Output Optimal Utility

Source: Russell & Norvig, AI - A Modern Approach

The Markov Property – True of Many Domains

- **Our policy at timepoint t is only dependent on the current state s**
  - $\pi(a|s) = P(A_t = a|S_t = s)$
- Although the agent has a history up until S_t
  - $\mathrm{H_t} = S_0, A_0, R_1 S_1, A_1, R_2 \ldots S_{t-1}, A_{t-1}, R_t, S_t$
- One may assume that all relevant information about the future is contained in the current state and action
  - $P(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a) = P(S_{t+1} = s', R_{t+1} = r|H_t = h_{t+1}, A_t = a)$
- This is a generalization of the Markov property to sequential decision problems
  - $P(S_{t+1}|S_t) = P(S_{t+1}|S_t, S_{t-1}, \ldots S_0)$

Source: Forest A.'s RL Course

RL with Finite States

Solving a Finite MDP

- **States:** A discrete and finite set $\mathcal{S}$
- **Actions:** A discrete and finite set $\mathcal{A}$
- **Transition Probabilities:** $P(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$
  - Defines the dynamics of the MDP
- The state-transition probabilities can be obtained from the transition probabilities
  - $p(s'|s,a) = \sum_{r \in \mathcal{R}} p(s', r|s, a)$ // Estimating state-transition by looking at reward of samples
- The **expected reward** can be obtained from the transition probabilities
  - $r(s, a) = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r|s, a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$
    // Estimating reward from transitions seen

Adapted from: Forest A.'s RL Course

Model-free RL: Q-learning

- Learning action-value functions
- Q(a,i): value of doing action a in state i
- Relationship between utility U of state and Q value
  - U(i) = max Q(a, i)
- Finding Q value based on whether transition probability is known
  - When M (transition is known)

$$Q(a,i) = R(i) + \sum_j M_{ij}^a \max_{a'} Q(a',j)$$

  - Estimating with TD method

$$Q(a, i) \leftarrow Q(a, i) + a\,(R(i) + \max_{a'} Q(a',j) - Q(a,i))$$

Source: Russell & Norvig, AI - A Modern Approach

RL with Deep Learning

- For small problems, like games, state-value function (U), action- utility value (Q), and transition functions (M), and policy functions are represented using a table

- But for large and realistic problems, number of states are countably large/ practically infinite
- Deep learning are excellent function approximators
  - Estimate Q-value i.e., action-value

- Not covered in this class

Exercise and Code – RL

- RL settings and solution methods

- Code: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l18-learning-agent/RL%20using%20Gym.ipynb

Source: Russell & Norvig, AI: A Modern Approach

Inverse Reinforcement Learning

- Given $\pi^*$and transition function M,
  - can we recover R

- Or, given execution traces corresponding to π^*
  - can we recover R?

- Applications
  - Path planning
  - Automated-driving

- Reference: Pieter Abbel's course slides: https://people.eecs.berkeley.edu/~pabbeel/cs287-fa12/slides/inverseRL.pdf

RL References

- Sutton and Barto's Book: http://incompleteideas.net/book/the-book.html
- Russell and Norvig, AI – A modern Approach
- David Silver's RL course, https://www.davidsilver.uk/teaching/

- Inverse RL
  - A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress, https://arxiv.org/abs/1806.06877, 2018
  - Pieter Abbel's course slides: https://people.eecs.berkeley.edu/~pabbeel/cs287-fa12/slides/inverseRL.pdf

RL Virtual School – Upcoming Event

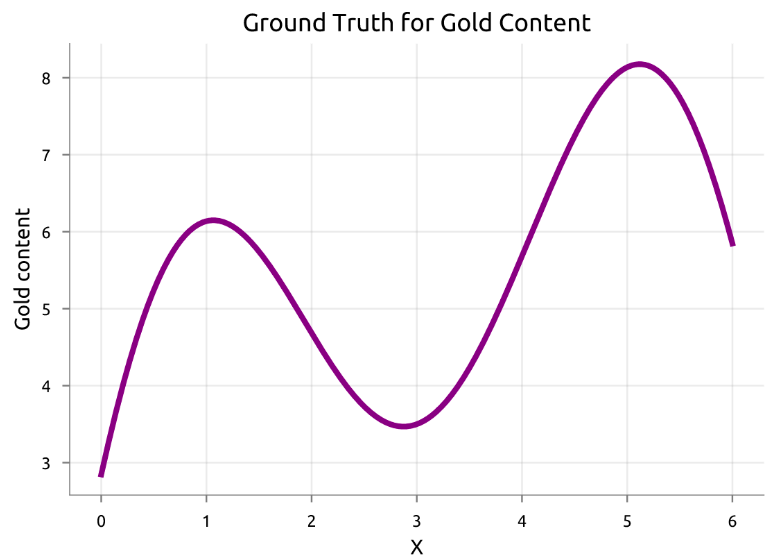
- When: March 25-26 and April 1,2,8,9
- Where: Virtual
- Details: https://rlvs.aniti.fr
  - https://rl-vs.github.io/rlvs2021

Schedule			
March 25th	9:00-9:10	Opening remarks	S. Gerchinovitz
	9:10-9:30	RLVS Overview	E. Rachelson
	9:30-12:30	RL fundamentals	E. Rachelson
	14:00-16:00	Deep Learning	D. Wilson
	16:30-17:30	Human behavioral agents	I. Rish
March 26th	10:00-12:00	Stochastic bandits	T. Lattimore
	14:00-16:00	Monte Carlo Tree Search	T. Lattimore
	16:30-17:30	Multi-armed bandits in clinical trials	D. A. Berry
April 1st	9:00-15:00	Deep Q-Networks and its variants	B. Piot
	15:15-16:15	Regularized MDPs	M. Geist
	16:30-17:30	TBA	M. Wang
April 2nd	9:00-12:30	Policy Gradients and Actor Critic methods	O. Sigaud
	14:00-15:00	Pitfalls in Policy Gradient methods	O. Sigaud
	15:30-17:30	Exploration in Deep RL	M. Pirotta
April 8th	9:00-11:00	Evolutionary Reinforcement Learning	D. Wilson, J.-B. Mouret
	11:30-12:30	TBA	S. Risi
	14:00-16:00	Micro-data Policy Search	K. Chatzilygeroudis, J.-B. Mouret
	16:30-17:30	TBA	
April 9th	9:00-13:00	RL tips and tricks	A. Raffin
	14:30-15:30	Symbolic representations and reinforcement learning	M. Garnelo
	15:45-16:45	Leveraging model-learning for extreme generalization	L. P. Kaelbling
	17:00-18:00	RLVS wrap-up	E. Rachelson

Bayesian Optimization (BO)

BO Problem Setting

- Input
  - A continuous variable: x $x \in R^d$
  - A function: f(x) $f: R^d \rightarrow R$
    - f is a black box for which detail, closed form or gradient is unknown
    - f is expensive to evaluate
    - Evaluations of f(x) = y may be noisy
- Output:
  - $x^* = \arg\min_x f(x)$



Source: Agnihotri & Batra, "Exploring Bayesian Optimization", Distill, 2020.

Details of Solving

- Acquisition function: heuristics about how desirable it is to evaluate a data point x_i, based on our present model
- Update by Bayes rules
  - At every step, a model of estimates and uncertainty at each point is updated using Bayes' rule

1. We first choose a surrogate model for modeling the true function f and define its **prior**.
2. Given the set of **observations** (function evaluations), use Bayes rule to obtain the **posterior**.
3. Use an acquisition function $\alpha(x)$, which is a function of the posterior, to decide the next sample point $x_t = \mathrm{argmax}_x \alpha(x)$.
4. Add newly sampled data to the set of **observations** and goto step #2 till convergence or budget elapses.

Acquisition function

$$x_{t+1} = argmax(\alpha_{PI}(x)) = argmax(P(f(x) \geq (f(x^+) + \epsilon)))$$

where,

$P(\cdot)$ indicates probability

ϵ is a small positive number

And, $x^+ = \mathrm{argmax}_{x_i \in x_{1:t}} f(x_i)$ where x_i is the location queried at i^{th} time step.

Recall: Bayes Theorem

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

H: Hypothesis, E: Evidence

BO Applications

- Hyperparameter tuning – Auto-AI
- Mining industry
- Sensor placement
- …

Exercise and Code

- Bayesian Optimization

  - Code: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l18-learning-agent/bayesian-optimization.ipynb

References: Bayesian Optimization

- Description:
  - https://distill.pub/2020/bayesian-optimization/
  - https://machinelearningmastery.com/what-is-bayesian-optimization/

- Papers
  - A Tutorial on Bayesian Optimization, Peter I. Frazier, https://arxiv.org/abs/1807.02811, 2018
  - Taking the Human Out of the Loop: A Review of Bayesian Optimization
    B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N.d. Freitas.
    Proceedings of the IEEE, Vol 104(1), pp. 148-175. 2016.
    DOI: 10.1109/JPROC.2015.2494218
  - A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning
    E. Brochu, V. M. Cora, N. De Freitas.
    CoRR, Vol abs/1012.2599. 2010.

- Scikit: https://scikit-optimize.github.io/stable/auto_examples/bayesian-optimization.html

Lecture 18: Concluding Comments

- We looked at a learning agent
- Reinforcement learning method
  - Various variations
- Bayesian Optimization

Concluding Segment

Upcoming Classes

15	**Mar 4 (Th)**	**Reasoning and Search**	**Semester - Midpoint**
16	Mar 9 (Tu)	Agent – Optimization	
17	Mar 11 (Th)	Agent – Handling Uncertain World	
18	Mar 16 (Tu)	Agent – Learning	
19	Mar 18 (Th)	Text: Data Prep (NLP)	Quiz 3
20	Mar 23 (Tu)	Text: Analysis - Supervised (NLP)_	
21	Mar 25 (Th)	Review, Paper presentations, Discussion	
22	Mar 30 (Tu)	Text: Advanced – Summarization, Sentiment	
23	Apr 1 (Th)	Text: Visualization, Explanation	
24	Apr 6 (Tu)	Paper presentations – Graduate students	Final assignment for Graduate students
25	Apr 8 (Th)	Case Study 1: Water (Structured + Text)	Quiz 4
26	Apr 13 (Tu)	Case Study 2: Finance (Structured+Text)	

Focus on Integrated Agent Behavior (Lectures 17, 18)

Paper Presentations – Graduate Students

- Select a paper appearing at a top-AI or data conference (AAAI, IJCAI, NeurIPS, SIGMOD, WWW, ICML, VLDB, …) during 2019-2020
- Present in class for 10 + 5 minutes of Q/A
- Things to cover
  - **Summary**: problem, solution, related work, contributions
  - **Opinion**: What you liked or did not like

About Next Lecture – Lecture 18

Lecture 18: Text Analysis

- What is text ?
  - Multi-lingual
- Nature of analysis possible
- How it complements numerical analysis
- Quiz 3