CS292F StatRL Lecture 6
RL Algorithms + Bandits

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Lecture 5

- Model-free RL that combines MC and DP
  - Temporal Difference methods and their variants
  - Function approximation
  - TD methods (TD prediction / policy evaluation) as one that minimizes the square Bellman error a semi-gradient update.
  - LSTD and its fast computation

- Policy gradients methods

Recap: TD-method in a nutshell --- attempting to simulate Bellman equations with roll-out data

TD-methods: what I did not cover

- n-step TD
- TD(λ) / Eligible traces
- Off-policy methods and the “deadly triad”
- Semi-gradient vs actual SGD?
- Projected Bellman error and Gradient-TD

(A large body of associated work on these.
Read Sutton and Barto Ch 7, CH 11-12 and the references therein.)

This lecture

- Continue with policy gradient methods
- Start exploration
  - Multi-armed bandits

Recap: Policy class and policy gradient methods

- Policy $\pi \in \Pi$
- Parametric policy class:

$$\Pi = \{\pi_\theta | \theta \in \mathbb{R}^d\}$$

- Goal: optimize the value
- Policy gradient methods
  - aim at learning the policy parameter by SGD.

How to estimate the gradient?

- Policy gradient theorem:

Proof of Policy Gradient Theorem

Proof of Policy Gradient Theorem

REINFORCE Algorithm

REINFORCE, A Monte-Carlo Policy-Gradient Method (episodic)

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$
Repeat forever:
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  For each step of the episode $t = 0, \ldots, T-1$:
    $G \leftarrow$ return from step t
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla_{\boldsymbol{\theta}} \ln \pi(A_t|S_t, \boldsymbol{\theta})$

Variance of the gradient estimate and convergence

- This is a non-convex optimization problem

- Convergence to a stationary point.
  - Choice of learning rate:
  - Then (Lemma 9.8 AJKS)

- Does it converge to the global optimal solution?
  - Yes, sometimes (under additional assumptions, see Ch 10 AJKS).

REINFORCE with a baseline

- REINFORCE with a given (arbitrary) baseline

- Choose it to approximate the value function.
- Why does it work?

Exercise: Calculate the variance when the baseline is chosen to be the correct value function.

Actor-Critic: Learn the baseline and use the baseline for "bootstrapping"

- **Actor:** The policy that performs actions.
- **Critic:** A value function approximation that evaluates the "actor"

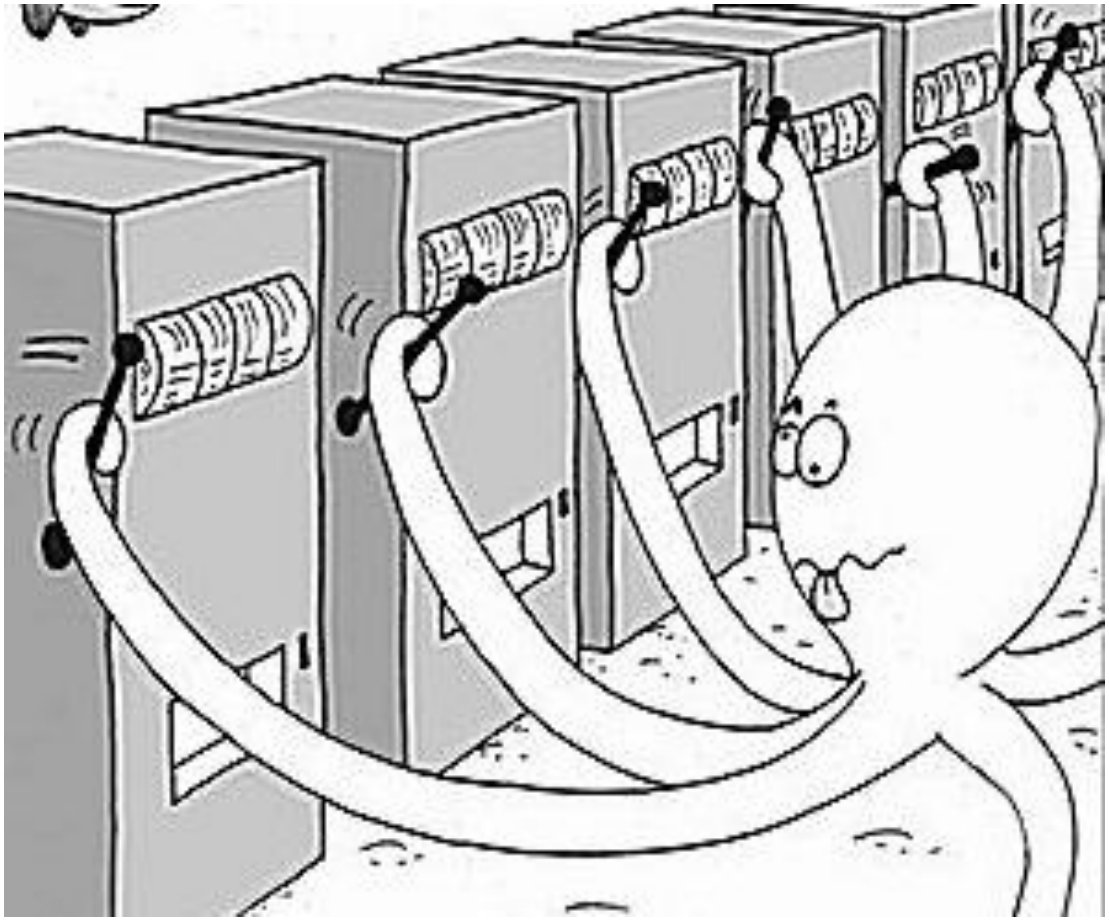
What I did not talk about

- Natural Policy Gradient
  - Extension of Natural Gradient (i.e., Fisher-Scoring)
  - Multiplicative form in updating the policy
  - Function approximation
- Global convergence of PG / NPG
- KL-divergence regularization / imitation learning / conservative policy iteration / trust-region policy iteration

Recap: Let us tackle different aspects of the RL problem one at a time

- Markov Decision Processes: (Lecture 1-3)
  - Dynamics are given no need to learn. planning only.
- RL algorithms: (Lecture 4-5)
  - Ideas on how to use “sampled” transitions / trials-and-errors to learn and plan at the same time (without touching upon exploration).
- **Strategic exploration: (Lecture 6- 9)**
  - Bandits: Explore-Exploit in simple settings
  - RL: Explore-Exploit in Learning MDPs

Slot machines and Multi-arm bandits



Multi-arm bandits: Problem setup

- No state. k-actions $a \in \mathcal{A} = \{1, 2, ..., k\}$
- You decide which arm to pull in every iteration

$$A_1, A_2, ..., A_T$$

- You collect a cumulative payoff of $\sum_{t=1}^{T} R_t$
- The goal of the agent is to maximize the expected payoff.
  - For future payoffs?
  - For the expected cumulative payoff?

How do we measure the performance of an **online learning agent**?

- The notion of “Regret”:
  - I wish I have done things differently.
  - Comparing to the best actions in the hindsight, how much worse did I do.

- For MAB, the regret is defined as follow

$$T \max_{a \in [k]} \mathbb{E}[R_t|a] - \sum_{t=1}^{T} \mathbb{E}_{a \sim \pi} \left[\mathbb{E}[R_t|a] \right]$$

Greedy strategy

- Expected reward

$$r(s,a) = Q(s,a) = \mathbb{E}\left[R_t \mid A_t = a\right].$$

- Estimate the expected reward

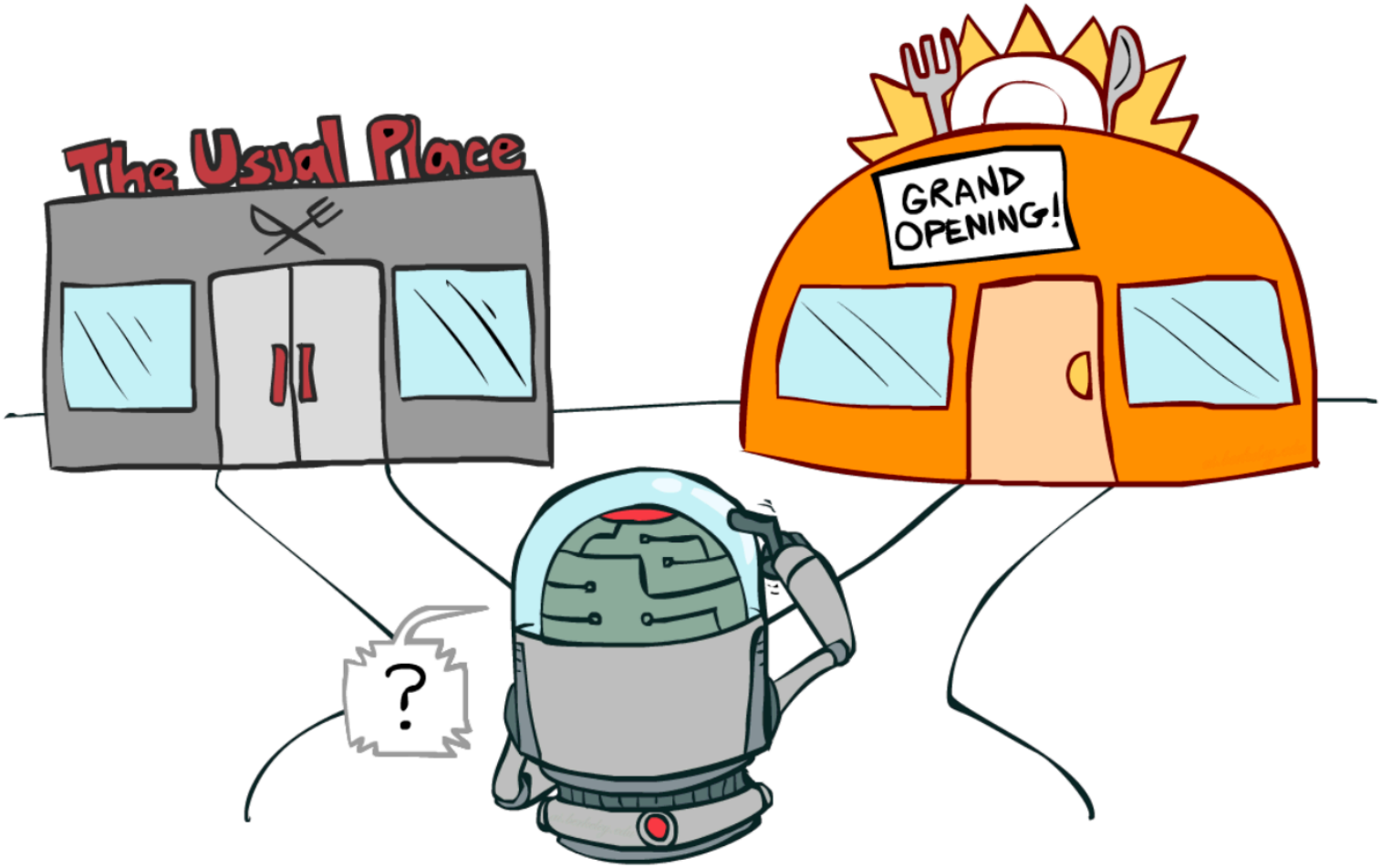
$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- Choose $A_t \doteq \arg\max_a Q_t(a),$

What is the issue with this strategy?

Exploration vs. Exploitation



(Illustration from Dan Klein and Pieter Abbeel's course in UC Berkeley)

Exploration first strategy

- Let's spend the first N step exploring.
  - Play each action for N / k times.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

- For t = N +1, N+2, …, T:

$$A_t \doteq \arg\max_a Q_t(a),$$

Recap: Concentration inequalities --- finite-sample bounds of LLN and CLT

- **Hoeffding's inequality**: Assume X_1, …, X_n are independent and their support bounded:

$$S_n = X_1 + \cdots + X_n$$

$$\mathrm{P}(S_n - \mathrm{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

- Easy version, if $0<X_i<B$, with probability $1-\delta$:

$$\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \leq \sqrt{\frac{B^2}{2n}\log(2/\delta)}$$

Regret analysis of Exploration First

Regret analysis of Exploration First

ε-Greedy strategy: one way to balance exploration and exploitation

- You choose with probability 1- ε

$$A_t \doteq \arg\max_a Q_t(a),$$

- With probability ε, choose an action uniformly at random!
  - Including the argmax.

- Carefully choose ε parameter.

A sketch of the analysis for ε-greedy

- In expectation, each arm is chosen for at least εt times.
- Condition on the number of times, apply Hoeffding's inequality / union bound for all t and a
- Regret bound is

$$\epsilon T \ + \sum_{t=1}^{T} C\sqrt{\frac{k}{\epsilon t}}$$

Optimism-in-the-face of uncertainty: Upper Confidence Bound algorithm

Martingale

- We say that a sequence of r.v. $X_1,\ldots,X_n,\ldots$ is a Martingale if for any n

$$\mathbf{E}(|X_n|) < \infty$$

$$\mathbf{E}(X_{n+1} \mid X_1, \ldots, X_n) = X_n.$$

- Example:
  - Random-walk: Total number of heads minus tails in n coin tosses

Azuma-Hoeffding's inequality

- **Azuma-Hoeffding's inequality**: Assume $X_1, \ldots, X_n$ are Martingale differences

$$S_n = X_1 + \ldots + X_n$$

$$\mathbb{P}\left[S_n \geq \epsilon\right] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

- Apply Azuma-Hoeffding's inequality to our problem

Regret analysis of UCB

Regret analysis of UCB

Summary of Exploration in Multi-Armed Bandits

- Explore-First
- eps-greedy
- UCB