Model Free Control

David Silver Lecture 5

Outline: 1. Introduction 2. On-policy Monte Carlo Control 3. On-policy Temporal Difference Learning 4. Off-policy Learning 5. Summary

Model Free Reinforcement Learning.

- Last lecture:
  - Model-Free prediction
  - Estimate the value function of an unknown MDP
- This lecture:
  - Model-free Control
  - Optimise the value function of an unknown MDP

Uses of Model-Free Control

Some example problems that can be modelled as MDPs

- ✓ Elevator
- ✓ Parallel Parking
- ✓ Ship Steering
- ✓ Bioreactor
- ✓ Helicopter
- ✓ Aeroplane logistics
- ✓ Quake
- ✓ Robocup Soccer
- ✓ Portfolio management
- ✓ Protein Folding
- ✓ Robot walking
- ✓ Game of Go.

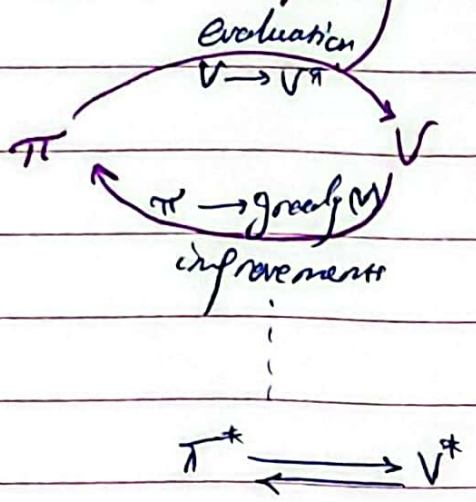
- For most of this problems, either:
  - MDP model is unknown, but experience can be sampled
  - MDP model is known, but is too big to use, except by sampling

Model Free Control can solve these problems.

On and off-policy Learning

- On-policy learning: 1) "Learn on the job"
  2) Learn about policy π from experience sampled from π
- Off-policy learning: 1) "look over someone's shoulder"
  2) Learn about policy π from experience sampled from μ

Generalized policy Iteration (Refresher)



Policy Evaluation Estimate V_π
  eg Iterative policy evaluation

Policy improvement Generate $\pi' \geq \pi$
  eg Greedy policy improvement

Generalized policy Iteration with Monte-Carlo Evaluation

Policy evaluation Monte-Carlo policy evaluation, $V = \cancel{V_\pi}$? → $Q = q_\pi$
needs dynamics MDP

Policy improvement $\underbrace{\epsilon\text{-Greedy policy improvement}}_{\text{do not have exploration}}$?

Model free policy Iteration Using Action-value Function

- Greedy policy improvement over $V(s)$ requires model of MDP

$$\pi'(s) = \arg\max_{a\in A} R_s^a + P_{ss'}^a V(s')$$

- Greedy policy improvement over $Q(s,a)$ is model-free

$$\pi'(s) = \arg\max_{a\in A} Q(s,a)$$

ε-Greedy Exploration

- Simplest idea for ensuring continual exploration
- All m actions are tried with non-zero probability
- with probability $1-\varepsilon$ choose the greedy action
- with probability ε choose an action at random

$$\pi(a|s) = \begin{cases} \varepsilon/m + 1 - \varepsilon & \text{if } a^* = \underset{a \in A}{\operatorname{argmax}}\, Q(s,a) \\ \varepsilon/m & \text{otherwise} \end{cases}$$

ε-Greedy Policy improvements

Theorem:

For any ε-greedy policy π, the ε-greedy policy π' with respect to q_π is an improvement, $V_{\pi'}(s) \geq V_\pi(s)$

$$q_\pi(s, \pi'(s)) = \sum_{a \in A} \pi'(a|s)\, q_\pi(s,a)$$

$$= \frac{\varepsilon}{m} \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \max_{a \in A} q_\pi(s,a)$$

$\pi(a|s) \geq \varepsilon/m$ زیرا

$$\geq \frac{\varepsilon}{m} \sum_{a \in A} q_\pi(s,a) + (1-\varepsilon) \sum_{a \in A} \frac{\pi(a|s) - \varepsilon/m}{1-\varepsilon}\, q_\pi(s,a)$$

$$= \sum_{a \in A} \pi(a|s)\, q_\pi(s,a) = V_\pi(s)$$

Therefore from policy improvement theorem $V_{\pi'}(s) \geq V_\pi(s)$

GLIE

Definition

Greedy in the Limit with Infinite Exploration (GLIE)

All state-action pairs are explored infinitely many times

$$\lim_{k \to \infty} N_k(s,a) = \infty$$

The policy converges on a greedy policy

$$\lim_{k \to \infty} \pi_k(a|s) = 1\left(a = \underset{a' \in A}{\arg\max}\, Q_k(s,a')\right)$$

For example, ε-greedy is GLIE if ε reduces to zero at $\varepsilon_k = \frac{1}{k}$

GLIE Monte-Carlo Control

Sample kth episode using π: $\{S_1, A_1, R_2, \ldots, S_T\} \sim \pi$

For each state S_t and action A_t in the episode,

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}\left(G_t - Q(S_t, A_t)\right)$$

Improve policy based on new action-value function

$$\varepsilon \leftarrow 1/k$$

$$\pi \leftarrow \varepsilon\text{-greedy}(Q)$$

Theorem:

GLIE Monte-Carlo Control converges to the optimal action-value function

$$Q(s,a) \longrightarrow q_*(s,a)$$

s.a.m

Off Policy Learning

- Evaluate target policy $\pi(a|s)$ to compute $v_\pi(s)$ or $q_\pi(s,a)$
- while following behaviour policy $\mu(a|s)$

$$\{S_1, A_1, R_2, \ldots, S_T\} \sim \mu$$

- Why is this important?
- Learn from observing humans or other agents
- Re use experience generated from old policies $\pi_1, \pi_2, \ldots, \pi_{t-1}$
- Learn about optimal policy while following exploratory policy
- Learn about multiple policies while following one policy

Importance Sampling

Estimate the expectation of a different distribution

$$\mathbb{E}_{X\sim P}[f(x)] = \sum P(X) f(X)$$

$$= \sum Q(x) \frac{P(x)}{Q(x)} f(x)$$

$$= \mathbb{E}_{X\sim Q}\left[\frac{P(x)}{Q(x)} f(x)\right]$$

for off policy Monte-Carlo

Use returns generated from μ to evaluate π

weight return G_t according to similarity between policies

Multiply importance sampling corrections along whole episode

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)\,\pi(A_{t+1}|S_{t+1})}{\mu(A_t|S_t)\,\mu(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_T|S_T)}{\mu(A_T|S_T)} G_t$$

Update value towards corrected return

$$V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^{\pi/\mu} - V(S_t)\right)$$

Cannot use if μ is zero when π is non-zero

Importance sampling ~~and~~ can dramatically increase ~~varance~~ variance

extremely high variance and for off policy you have to use TD

s.a.m