TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Deep Reinforcement Learning

Definition of Reinforcement Learning

RL is defined by the following properties:

- An environment with **state**.
- State changes are influenced by **sequential decisions**.
- Reward (or loss) depends on **making decisions** that lead to **desirable states**.

Reinforcement Learning Examples

- Board games (chess or go)
- Atari Games (pong)
- Robot control (driving)
- Dialog
- Life

Policies

A policy is a way of behaving.

Formally, a (nondeterministic) policy maps a state to a probability distribution over actions.

$\pi(a_t|s_t)$ probability of action a_t in state s_t

Imitation Learning

Construct a training set of state-action pairs (s, a) from experts.

Define stochastic policy $\pi_\Phi(s)$.

$$\Phi^* = \operatorname*{argmin}_\Phi E_{(s,a)\sim\text{Train}} \; -\ln \pi_\Phi(a \mid s)$$

This is just cross-entropy loss where we think of a as a "label" for s.

Dangers of Imperfect Imitation Learning

Perfect imitation learning would reproduce expert behavior. Imitation learning is **off-policy** — the state distribution in the training data is different from that defined by the policy being learned.

Imitating experts, such as expert fire eaters, can be dangersous. "Don't try this at home".

Also, it is difficult to exceed expert performance by imitating experts. But this can happen.

Markov Decision Processes (MDPs)

An MDP consists of a set $\mathcal{S}$ of states, a set $\mathcal{A}$ of allowed actions, a reward function R and a next-state probability function P_T. We will use the following notation.

$s_t \in \mathcal{S}$ is the state at time t

$a_t \in \mathcal{A}$ is the action taken at time t.

$r_t = R(s_t, a_t) \in \mathbb{R}$ is the reward at time t

$P_T(s_{t+1}|s_t, a_t)$ is the probability of s_{t+1} given s_t and a_t.

The function $R(s, a)$ can allow for a cost of the action a.

Optimizing Reward

In RL we maximize reward rather than minimize loss.

$$\pi^* = \underset{\pi}{\text{argmax}}\ R(\pi)$$

$$R(\pi) = E_\pi \sum_{t=0}^{T} r_t \qquad \text{episodic reward (go)}$$

$$\text{or } E_\pi \sum_{t=0}^{\infty} \gamma^t r_t \qquad \text{discounted reward (financial planning)}$$

$$\text{or } \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} r_t \qquad \text{asymptotic average reward (driving)}$$

The Value Function

For discounted reward:

$$V^{\pi}(s) = E_{\pi} \sum_{t} \gamma^{t} r_t \mid \pi, \ s_0 = s$$

$$V^{*}(s) = \sup_{\pi} V^{\pi}(s)$$

$$\pi^{*}(a|s) = \operatorname*{argmax}_{a} R(s,a) + \gamma E_{s' \sim P_T(s'|s,a)} V^{*}(s')$$

$$V^{*}(s) = \max_{a} R(s,a) + \gamma E_{s' \sim P_T(\cdot|s,a)} V^{*}(s')$$

Value Iteration

Suppose the state space and action space are finite.

In that case we can do value iteration.

$$V_0(s) = 0$$

$$V_{i+1}(s) = \max_a \; R(s,a) + \gamma E_{s' \sim P_T(\cdot|s,a)} \, V_i(s')$$

If all rewards are non-negative then

$$V_{i+1}(s) \geq V_i(s) \quad V_i(s) \leq V^*(s) \quad \text{so} \lim_{i \to \infty} \; V_i(s) \text{ exists}$$

Value Iteration

Theorem: **For discounted reward**

$$V_\infty(s) \doteq \lim_{i \to \infty} V_i(s) = V^*(s)$$

Proof

$$
\begin{aligned}
\Delta &\doteq \max_s \; V^*(s) - V_\infty(s) \\
&= \max_s \begin{pmatrix} \max_a R(s,a) + E_{s'|a}\gamma V^*(s') \\ -\max_a R(s,a) + E_{s'|a}\gamma V_\infty(s') \end{pmatrix} \\
&\leq \max_s \max_a \begin{pmatrix} R(s,a) + E_{s'|a}\gamma V^*(s') \\ -R(s,a) + E_{s'|a}\gamma V_\infty(s') \end{pmatrix} \\
&= \max_s \max_a \; E_{s'|a} \; \gamma(V^*(s') - V_\infty(s)) \\
&\leq \gamma\Delta
\end{aligned}
$$

The Q Function

For discounted reward:

$$Q^{\pi}(s,a) = E_{\pi} \sum_{t} \gamma^{t} r_t \mid \pi,\ s_0 = s,\ a_0 = a$$

$$Q^{*}(s,a) = \sup_{\pi} Q^{\pi}(s,a)$$

$$\pi^{*}(a|s) = \operatorname*{argmax}_{a} Q^{*}(s,a)$$

$$Q^{*}(s,a) = R(s,a) + \gamma E_{s' \sim P_T(\cdot|s,a)} \max_{a'} Q^{*}(s',a')$$

Q Function Iteration

It is possible to define Q-iteration by analogy with value iteration, but this is generally not discussed.

Value iteration is typically done for finite state spaces. Let S be the number of states and A be the number of actions.

One update of a Q table takes $O(S^2A^2)$ time while one update of value iteration is $O(S^2A)$.

Q-Learning

When learning by updating the Q function we typically assume a parameterized Q function $Q_\Phi(s, a)$.

Bellman Error:

$$\text{Bell}_\Phi(s,a) \doteq \left(Q_\Phi(s,a) - \left(R(s,a) + \gamma \; E_{s' \sim P_T(s'|s,a)} \; \max_{a'} \; Q_\Phi(s',a') \right) \right)^2$$

Theorem: If $\text{Bell}_\Phi(s,a) = 0$ for all (s,a) then the induced policy is optimal.

Algorithm: Generate pairs (s,a) from the policy $\text{argmax}_a \; Q_\Phi(s_t,a)$ and repeat

$$\Phi \; \texttt{-=} \; \eta \nabla_\Phi \; \text{Bell}_\Phi(s,a)$$

Issues with Q-Learning

Problem 1: Nearby states in the same run are highly correlated. This increases the variance of the cumulative gradient updates.

Problem 2: SGD on Bellman error tends to be unstable. Failure of Q_Φ to model unused actions leads to policy change (exploration). But this causes Q_Φ to stop modeling the previous actions which causes the policy to change back ...

To address these problems we can use a **replay buffer**.

Using a Replay Buffer

We use a replay buffer of tuples (s_t, a_t, r_t, s_{t+1}).

Repeat:

1. Run the policy $\text{argmax}_a\, Q_\Phi(s, a)$ to add tuples to the replay buffer. Remove oldest tuples to maintain a maximum buffer size.
2. $\Psi = \Phi$
3. for N times select a random element of the replay buffer and do

$$\Phi \text{ -= } \eta \nabla_\Phi \left(Q_\Phi(s_t, a_t) - \left(r_t + \gamma \max_a Q_\Psi(s_{t+1}, a)\right)^2\right.$$

Replay is Off-Policy

Note that the replay buffer is from a **mixture of policies** and is **off-policy** for $\text{argmax}_a \; Q_\Phi(s, a)$. This seems to be important for stability.

This seems related to the issue of stochastic vs. deterministic policies. More on this later.

Multi-Step Q-learning

$$\Phi \mathrel{-}= \sum_t \nabla_\Phi \left(Q_\Phi(s_t, a_t) - \sum_{\delta=0}^{D} \gamma^\delta r_{(t+\delta)} \right)^2$$

Human-level control through deep RL (DQN)

Mnih et al., Nature, 2015. (Deep Mind)

We consider a CNN $Q_\Phi(s, a)$.

Watch The Video

https://www.youtube.com/watch?v=V1eYniJ0Rnk

Asynchronous Q-Learning (Simplified)

No replay buffer. Many asynchronous threads each repeating:

$\tilde{\Phi} = \Phi$ (retrieve Φ)

using policy $\text{argmax}_a Q_{\tilde{\Phi}}(s, a)$ compute

$$s_t, a_t, r_t, \ldots, s_{t+K}, a_{t+K}, r_{t+K}$$

Φ -= $\eta \sum_{i=t}^{t+K-D} \nabla_{\tilde{\Phi}} \left(Q_{\tilde{\Phi}}(s_i, a_i) - \sum_{\delta=0}^{D} \gamma^\delta r_{i+\delta} \right)^2$ (update Φ)

END