

The background of the slide is an abstract, isometric pattern of numerous 3D cubes. The cubes are arranged in a dense, overlapping field that recedes into the distance. The color palette is composed of various shades of blue, ranging from a deep, saturated blue to a very light, almost white blue, creating a sense of depth and texture.


Deep Reinforcement Learning

372.2.5910

Ben-Gurion University of the Negev

Lecture Notes

WRITTEN BY: Hadar Sharvit

ALSO AVAILABLE ON GITHUB 

CONTACT ME AT: Hadar.Sharvit1@mail.huji.ac.il

BASED ON: Lectures given by Gilad Katz

CHAPTERS & BOOK COVER BY: Rohit Choudhari/Unsplash



Contents

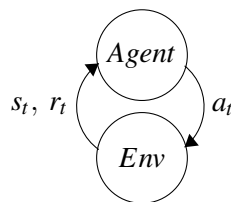
1	Hello world	5
1.1	Terminology	5
1.1.1	State & Observation	5
1.1.2	Action spaces	6
1.1.3	Policy	6
1.1.4	Trajectories	6
1.1.5	Reward	7
1.1.6	The goal of RL	7
1.1.7	Value function	7
1.1.8	The optimal Q-Function and the optimal action	8
1.1.9	Bellman Equations	8
1.1.10	Advantage function	9
1.2	Kinds of RL Algorithms	9
1.2.1	Model-Free vs Model-Based RL	9
1.2.2	What do we learn in RL	9
1.3	Intro to policy optimization	10
2	RL basics	13
2.1	Motivation	13
2.2	When to use RL?	14
2.3	Markov Decision Processes (MDP)	14
2.3.1	The Markov Property	14
2.4	Goals and Rewards	16

2.5	Policies, Value-function and Q-function	16
2.6	The Bellman equation	16
2.7	Policy Iteration	18
2.8	Value iteration	18
2.9	Monte-Carlo	18
2.9.1	Approximating Value-function	20
2.9.2	Approximating policies	20
2.10	On/Off-Policy methods	20
2.10.1	Importance sampling for Off-Policy methods	21
2.11	ϵ-Greedy Algorithms	22
2.12	Temporal Difference (TD) Learning	22
2.12.1	On-Policy TD Control: SARSA	23
2.12.2	Off-Policy TD Control: Q-Learning	24
3	DQN & it's derivatives	25
3.1	Deep Q-Network	26
3.1.1	Architecture	26
3.1.2	Training DQN	26
3.2	Double Deep Q-Network (DDQN)	27
3.3	Dueling network	28
3.3.1	Implementation	29
4	Policy Gradients	31
4.1	The Policy-Gradient theorem	31
4.2	The REINFORCE Algorithm	32
4.2.1	REINFORCE with Baseline	33

1. Hello world

R This chapter is provided as a preliminary, and is not part of the course. It is based on OpenAI's Spinning Up docs (For further references see [here](#)).

Reinforcement Learning (RL) is the study of agents and how they learn by trial and error. The two main components of RL are the *agent* and the *environment* - The agent interacts with the environment (also known as taking a "step") by seeing a (sometimes partial) *observation* of the environment's *state*, and then decides which *action* should be. The agent also perceives a reward from the environment, which is essentially a number that tells the agent how good the state of the world is, and the agent's goal is to maximize the *cumulative reward*, called *return*.



1.1 Terminology

let's introduce some additional terminology

1.1.1 State & Observation

The complete description of the environment/world's state is the *state* s , and an *observation* o is a partial description of s . We usually work with an observation, but wrongly denote it as s - we are going to stick with this convention.

R if $o = s$ we say that the environment is *fully observed*. Otherwise, it is *partially observed*.

1.1.2 Action spaces

Is the set of all valid actions in the environment. The action space could be discrete (like the action to move in one of the direction $\{\uparrow, \downarrow, \leftarrow, \rightarrow\}$) or continuous (like the action to move the motor with $\alpha \in \mathbb{R}$ Newtons of force)

1.1.3 Policy

Is a set of rules used by our agent to decide on the next action. It can be either deterministic or stochastic $a_t \sim \pi(\cdot|s_t)$. Under the scope of deep RL, the policy is a parameterized function, i.e it is a mapping with parameters θ that should be learned in some optimization process. A deterministic Policy could be implemented, for example, using some basic MLP architecture. For a stochastic policy, the two most common types are *Categorical* policy (for discrete action space) and *Diagonal-Gaussian* policy (for continuous action space)

Categorical (stochastic) Policy

Is essentially a classifier, mapping discrete states to discrete actions. For example, you could build a basic NN that takes in the observation and outputs action probabilities (after applying softmax). Denoting the last layer as $P_\theta(s)$, we can treat the actions as indices so the log-likelihood for action a is

$$\log \pi_\theta(a|s) = \log [P_\theta(s)]_a \quad (1.1)$$

Given $P_\theta(s)$, we can also sample from the distribution (one can use PyTorch Categorical to sample from a probability vector)

Diagonal-Gaussian (stochastic) Policy

Is a policy that can be implemented using a neural network that maps observations to *mean* actions, under the assumption that the action probability space can be represented by some multivariate Gaussian with diagonal covariance matrix, which can be represented in two ways

- we use $\log \text{diag}(\Sigma) = \log \sigma$ which is *not* a function of the state s (σ is a vector of standalone parameters)
- we use a NN that maps from $s \rightarrow \log \sigma_\theta(s)$

we use $\log \sigma$ and not σ as the log takes any value $\in (-\infty, \infty)$, unlike σ that only takes values in $[0, \infty)$, making it harder to train.


Once the mean action $\mu_\theta(s)$ and the std $\sigma_\theta(s)$ are obtain, the action is sampled as $a = \mu_\theta(s) + \sigma_\theta(s) \otimes z$, where $z \sim N(0, 1)$ and \otimes is element-wise multiplication (This is similar to VAEs).

The log-likelihood of a k -dimensional action $a \in \mathbb{R}^k$ for a diagonal-Gaussian with mean μ_θ and std σ_θ can be simplified if we remember that when Σ is diagonal, a k -multivariate Gaussian's PDF is equivalent to the product of k one-dimension Gaussian PDF, hence

$$\log [\pi_\theta(a|s)] = \log \left[\frac{\exp \left[-\frac{1}{2}(a - \mu)^T \Sigma^{-1}(a - \mu) \right]}{\sqrt{(2\pi)^k |\Sigma|}} \right] = \dots = -\frac{1}{2} \left[\sum_{i=1}^k \left(\frac{(a_i - \mu_i)^2}{\sigma_i^2} + 2 \log \sigma_i \right) + k \log 2\pi \right] \quad (1.2)$$

1.1.4 Trajectories

we denote the trajectory as $\tau = (s_0, a_0, s_1, a_1, \dots)$ where the first state s_0 is randomly sampled from some start-state distribution $s_0 \sim \rho_0$. A new state is obtained from the previous state and action in either a stochastic or deterministic process.

 τ is also noted as "episode" or "rollout"

1.1.5 Reward

The reward r_t is some function of our states and action, and the goal of the agent is to maximize the cumulative reward over some trajectory τ .

- Finite-horizon un-discounted return: $R(\tau) = \sum_{t=0}^T r_t$
- Infinite-horizon discounted return: $R(\tau) = \sum_{t=0}^{\infty} \gamma^t r_t, \gamma \in (0, 1)$

Not adding a converging term γ^t means that our infinite sum may diverge, but also it manifests the concept of "reward now > reward later"

1.1.6 The goal of RL

We always wish to find a policy π^* which maximizes the expected return when the agent acts according to it. Under the assumption of stochastic environment and policy, we can write the probability to obtain some trajectory τ of size T , given a policy π as

$$P(\tau|\pi) = \rho_0(s_0) \prod_{t=0}^{T-1} \underbrace{P(s_{t+1}|s_t, a_t)}_{\text{Pr. to reach } s_{t+1} \text{ from } s_t \text{ when applying } a_t.} \cdot \underbrace{\pi(a_t|s_t)}_{\text{Pr. to choose action } a_t \text{ when in state } s_t.} \quad (1.3)$$

The expected return is by definition the sum of returns given all possible trajectories, weighted by their probabilities

$$J(\pi) = \int_{\tau} P(\tau|\pi) R(\tau) = \mathbb{E}_{\tau \sim \pi} [R(\tau)] \quad (1.4)$$

and w.r.t to this objective, we wish to find

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}} J(\pi) \quad (1.5)$$

1.1.7 Value function

We can think of the expected return given some specific state, or some specific state-action pair as the "value" of the state, or the state-action pair. Those are simply the expected return conditioned with some initial state or action

- On-policy Value function $V^{\pi}(s)$: if you start from s and act according to π , the expected reward is

$$V^{\pi}(s) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s]$$

- On-policy Action-Value function $Q^{\pi}(s)$: if you start from s , take an action a (which may or may not come from π) and only then act according to π , the expected reward is

$$Q^{\pi}(s, a) = \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a]$$

when finding a value function or an action-value function that maximizes the expected reward, we scan various policies and extract $V^*(s) = \max_{\pi \in \Pi} V^\pi(s)$ or $Q^*(s, a) = \max_{\pi \in \Pi} Q^\pi(s, a)$.

We can also find a relation between V and Q :

$$\begin{aligned}
 V^\pi(s) &= \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s] \\
 &= \sum_{\tau \sim \pi} Pr[R(\tau) | s_0 = s] R(\tau) \\
 &= \sum_{\tau \sim \pi} \sum_{a \sim \pi} Pr[R(\tau), a | s_0 = s] R(\tau) \text{ [Total prob.]} \\
 &= \sum_{a \sim \pi} Pr[a | s_0 = s] \sum_{\tau \sim \pi} Pr[R(\tau) | s_0 = s, a_0 = a] R(\tau) \\
 &= \sum_{a \sim \pi} Pr[a | s_0 = s] \mathbb{E}_{\tau \sim \pi} [R(\tau) | s_0 = s, a_0 = a] \\
 &= \sum_{a \sim \pi} Pr[a | s_0 = s] Q^\pi(s, a) \\
 &= \mathbb{E}_{a \sim \pi} Q^\pi(s, a)
 \end{aligned} \tag{1.6}$$

where in the 4'th line we used the fact that the probability of both $R(\tau)$ and a is the same as summing over all possible a and conditioning the probability $Pr[R(\tau)]$ given a . In terms of optimality, notice that as $V^*(s)$ is the optimal value function for a specific s , and for any a , and $Q^*(s, a)$ is the optimal value for a specific s and a , taking $\max Q$ over all a is exactly $V(s)$. Specifically

$$V^*(s) = \max_a Q^*(s, a) \tag{1.7}$$

1.1.8 The optimal Q-Function and the optimal action

$Q^*(s, a)$ gives the expected return for starting in s and taking action a , and then acting according to the optimal policy. As the optimal policy will select, when in s , the action that maximizes the expected return for when the initial state is s , we can obtain the optimal action a^* by simply maximizing over all values of Q^*

$$a^*(s) = \operatorname{argmax}_a Q^*(s, a) \tag{1.8}$$

We also note that if there are many optimal actions, we may choose one randomly

1.1.9 Bellman Equations

An important idea for all value functions is that the value of your starting point is the reward you expected to get from being there + the value of wherever you land next

$$V^\pi(s) = \mathbb{E}_{a \sim \pi, s' \sim P} [r(s, a) + \gamma V^\pi(s')] \tag{1.9}$$

$$Q^\pi(s, a) = \mathbb{E}_{s' \sim P} [r(s, a) + \gamma \mathbb{E}_{a' \sim \pi} [Q^\pi(s', a')]] \tag{1.10}$$

and optimality is obtained for

$$V^*(s) = \max_{a \sim \pi} \mathbb{E}_{s' \sim P} [r(s, a) + \gamma V^*(s')] \tag{1.11}$$

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[r(s, a) + \gamma \max_{a' \sim \pi} [Q^*(s', a')] \right] \tag{1.12}$$

1.1.10 Advantage function

Sometimes we only care if an action is better than others on average, and do not care as much for its' value on its own. The advantage function $A^\pi(s, a)$ describes how better is taking action a (that can be from π or not) when in s compared to selecting some random action $a' \sim \pi$, assuming you act according to π afterwards.

$$\begin{aligned} A^\pi(s, a) &= Q^\pi(s, a) - V^\pi(s) \\ &= Q^\pi(s, a) - \mathbb{E}_{a \sim \pi}[Q^\pi(s, a)] \end{aligned} \quad (1.13)$$

By calculating the advantage, we ask if the Q function of some candidate action a (given some state s) is larger then the average Q -function associated with the examination of all other actions taken by our policy.

1.2 Kinds of RL Algorithms

1.2.1 Model-Free vs Model-Based RL

In some cases we have a closed form of how our environment behaves. For example, we may know the probability space $P[s'|s, a]$, i.e we know that probability to transition from some s to some other s' given arbitrary action a . It may also be the case that our model can be described using some equation of motion. Either way, we can use this knowledge to formulate an optimal solution, which in many cases translate to some greedy approach of scanning various states and thinking ahead.

Some problems are that such model may not even be available (or even known), as for example, I do not have a model of how my chess opponent may play. Furthermore, greedy approaches usually mean brute-forcing all possible solutions.

In Model-Free RL, the model is not available, and we are trying to understand how the environment behaves by exploring it in some non-exhaustive manner. This means that model-free RL are likely to be not sample-efficient, though they are usually easier to implement

1.2.2 What do we learn in RL

After going through the taxonomy, we can ask whether we wish to learn the Q -function, the value-function, the policy or the environment model itself.

Under model-free RL

- Policy optimization: we parameterize the policy $\pi_\theta(a|s)$ and find optimum w.r.t the return $J(\pi_\theta)$. Such optimization is usually *on-policy*, meaning that the data used in the training process is only data given while acting according to the most recent version of the policy. In policy optimization we also find an approximator value function $V_\phi(s) \approx V^\pi(s)$. Some examples are *A2C, A3C, PPO*.
- Q-Learning: approximate $Q_\theta(s, a) \approx Q^*(s, a)$. Usually the objective is some form of the bellman equation. Q-Learning is usually *off-policy*, meaning that we use data from any point during training. Some examples are *DQN, C51*.

Compared to Q-Learning, that approximates Q^* , policy optimization finds exactly what we wish for - how to act optimally in the environment. Also, there are models that combine the two approaches, as *DDPG* for example, which learns both a Q function and an optimal policy.

Under model-based RL

cannot be clustered as easily, though some of the (many) approaches include methods of planning techniques to select actions that are optimal w.r.t to the model.

1.3 Intro to policy optimization

We aim to maximize the expected return $J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)]$, and we assume the finite-horizon undiscounted return (∞ -horizon is nearly identical). Our goal is to optimize π_θ with a gradient step

$$\theta_{k+1} = \theta_k + \alpha \underbrace{\nabla_{\theta} J(\pi_{\theta})}_{\text{Policy gradient}} \big|_{\theta_k} \quad (1.14)$$

and to do so, we must find a numerical expression for the policy gradient. As $J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau)] = \int_{\tau} P(\tau|\theta)R(\tau)$, we might as well write down a term for the probability of a trajectory

$$P(\tau|\theta) = \rho_0(s_0) \prod_{t=0}^T P(s_{t+1}|s_t, a_t) \pi_\theta(a_t|s_t) \quad (1.15)$$

Using the log-derivative trick, $\frac{d}{dx} \log x = \frac{1}{x}$, meaning that $x \frac{d \log x}{dx} = 1$. rewrite 1 as $\frac{d}{dx} x$, Substitute $x \leftrightarrow P(\tau|\theta)$ and $\frac{d}{dx} \leftrightarrow \nabla_{\theta}$ and we have that $P(\tau|\theta) \nabla_{\theta} \log P(\tau|\theta) = \nabla_{\theta} P(\tau|\theta)$. We will use this later. Now, lets expand the log term

$$\log P(\tau|\theta) = \log \rho_0(s_0) + \sum_{t=0}^T [\log P(s_{t+1}|s_t, a_t) + \log \pi_\theta(a_t|s_t)] \quad (1.16)$$

When deriving w.r.t θ , we are only left with the last term (the others only depend on the environment and not our agent), hence

$$\nabla_{\theta} \log P(\tau|\theta) = \nabla_{\theta} \sum_{t=0}^T \log \pi_\theta(a_t|s_t) = \sum_{t=0}^T \nabla_{\theta} \log \pi_\theta(a_t|s_t) \quad (1.17)$$

Notice the use of linearity in the second transition. Consequently, we re-write the expected return using eq 1.17 -

$$\begin{aligned} \nabla_{\theta} J(\pi_{\theta}) &= \nabla_{\theta} \int_{\tau} P(\tau|\theta) R(\tau) \\ &= \int_{\tau} \nabla_{\theta} P(\tau|\theta) R(\tau) \\ &= \int_{\tau} P(\tau|\theta) \nabla_{\theta} \log P(\tau|\theta) R(\tau) \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} [\nabla_{\theta} \log P(\tau|\theta) R(\tau)] \\ &= \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t|s_t) R(\tau) \right] \end{aligned} \quad (1.18)$$

In the 3rd transition we used the log-derivative trick, and in the last transition we used the expression from 1.17.

The last term is an expectation, hence can be estimated using mean - given a collected set $D =$

$\{\tau_1, \tau_2, \dots, \tau_N\}$ of trajectories obtained by letting our agent act in the environment using π_θ we can write

$$\nabla_\theta J(\pi_\theta) \approx \frac{1}{|D|} \sum_{\tau \in D} \sum_{t=0}^T \nabla_\theta \log \pi_\theta(a_t | s_t) R(\tau) \quad (1.19)$$

- R** It should be stated that this "Loss" term is not really a "loss" like we know from supervised learning. First of all, it does not depend on a fixed data distribution - here, the data is sampled from the recent policy. More importantly, it does not measure performance! the only thing it makes sure of is that given the *current* parameters, it has the negative gradient of performance. After this first step of gradient descent, there is no more connection to performance. This means that the loss minimization has no guarantee to improve expected return. This should come as a warning to when we look at the loss going down thinking that all is well - in policy gradients, this intuition is wrong, and we should only look at the average return.

2. RL basics

2.1 Motivation

Current malware detection platforms often deploy an ensemble of detectors to increase overall performance. This approach creates lots of redundancy, as in most cases one detector is enough, and it is of course computationally expensive and time consuming, compared to one detector.

We can come up with a simple improvement - query a subset of detectors and decide based on their classification if more detectors are needed. If we observe our approach under the scope of classification, it may very well be the case that training a model w.r.t to every set of detectors is needed, as we cannot evaluate the performance of a subset of detectors without actually learning how they performed. As this is computationally hard for large detectors, this is not a preferred approach. Instead, we can use RL:

Suppose we use four detectors, and our agent takes as input the vector $[-1, -1, -1, -1] \in \mathbb{R}^4$, which is considered an initial state. The agent will choose a set of detectors/detector configurations, and a classification measurement of either "*malicious*" or "*benign*" will be taken. The decisions of the agent will be based on a reward mechanism that takes uses the values of TP, FP (correctly classified the content as "*malicious*" or "*benign*") and FN, FN (incorrectly classified to "*malicious*" or "*benign*"). We will "punish" using $C(t)$, which is a function that depends on the time it took for the detectors to run. We can see that regardless of how many detectors were used, if we are right - the

Exp. #	TP	TN	FP	FN
1	1	1	$-C(t)$	$-C(t)$
2	10	10	$-C(t)$	$-C(t)$
3	100	100	$-C(t)$	$-C(t)$

Table 2.1: Three suggested reward mechanisms for a malware detection platform

reward is constant (experiment with 1, 10, 100). On the other hand, if we were wrong, we subtract

$C(t)$ which increases with the time t that has passed. As it is now "painful" to use more detectors, the reward incentivises our model to only use more detectors if that addition translated to higher success rates. As our model efficiently scans through the state space, it is able to outperform (at least conceptually) the suggested "check-all" classification approach that was previously introduced.

2.2 When to use RL?

Not all cases adhere to the framework of RL. Here are some rules

- our data should be in the form of trajectories - a set of distinguishable states $s_0 \rightarrow s_1 \rightarrow \dots \rightarrow s_N$
- need to make a sequence of related decisions - if every decision is independent, like classifying 4 images of cats-vs-dogs, don't use RL.
- the actions we perform result in some feedback - either positive or negative
- Tasks that are more suitable include both learning and planning - we learn our environment and plan an optimal behaviour
- the data is not provided apriori, and its' distribution changes with our action choice. This means that we must make sure our agent effectively explores the entire data, and does not settle in some small subspace.

2.3 Markov Decision Processes (MDP)

consists of the following

- States: that make up the environment. could be either discrete or continuous.
- Actions: by taking an action we transition from one state to another. In a deterministic process, we have that $P(s'|s, a) = 1$
- Reward: taking action $a \in A$ from state $s \in S$ results in reward $R(s, a) \in \mathbb{R}$

In finite MDP, s, a, r are all finite.

2.3.1 The Markov Property

the distribution over the future states depends only on the present state and action

$$Pr[s_{t+1}|s_1, a_1, s_2, a_2, \dots, s_t, a_t] = Pr[s_{t+1}|s_t, a_t] \quad (2.1)$$

In poker, for example, the markovian property does not hold as a player's current hand depends on the actions and hands of previous hands and/or players. A traffic light, on the other hand, is completely markovian as it is based on deterministic rules.

Using the markovian property, we can define the probability to reach a certain state s' with a certain reward r , as

$$Pr[s', r|s, a] \equiv Pr[s_{t+1} = s', r_{t+1} = r|s_t = s, a_t = a] \quad (2.2)$$

The probabilities induced by all event in S and R make up a probability space, hence

$$\forall s \in S \forall a \in A : \sum_{s' \in S} \sum_{r \in R} Pr[s', r|s, a] = 1 \quad (2.3)$$

The expected reward for state-action pairs, namely, what should we anticipate (in terms of reward) when performing the action a from the state s is

$$r(s, a) \equiv \mathbb{E}[r_{t+1}|s_t = s, a_t = a] = \sum_{r \in R} r \sum_{s' \in S} Pr[s', r|s, a] \quad (2.4)$$

where notice that the probability for a specific reward is the sum over all states, given the specific r (hence the sum over $s' \in S$).

We can also phrase our reward in terms of state-action-next state triplets, namely, what should we anticipate (in terms of reward) when performing the action a that takes us from state s to state s'

$$r(s, a, s') \equiv \mathbb{E}[r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'] = \sum_{r \in R} r \frac{Pr[s', r | s, a]}{Pr[s' | s, a]} \quad (2.5)$$

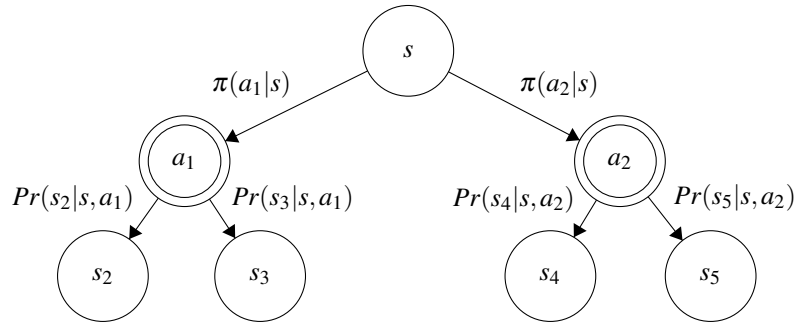
Where we can think of the probability fraction as the number of events that reach s' (from s after performing a) and provide reward r , out of all the event that reach s' (from s after performing a) given any reward.

MDPs are very flexible

- both states and actions could be either abstract ($s = \text{"sad"}$, $s = \text{"happy"}$, $a = \text{"take a nap"}$) or well-defined (like $s = \text{sensor readings}$, or $a = \text{turn on a switch}$).
- the time intervals may not be constant (some transitions are slow while other are fast)
- the setting of an MDP does not need to be an exact copy of the real-world model. For example, a set of sensors may be enough to describe a robotic arm, even though there are many more aspects that the arm is made up of (that are not as relevant).



At this point it may be helpful to look at what is known as the "*Backup Diagram*", That describes how the states are propagated based on the actions chosen by π and the probabilities induces by the environment.



We can write, for example, the probability to transition from s to s_5 by performing an action a_s as $P(s_5 | s, a_2) = \pi(a_2 | s) Pr(s_5 | s, a_2)$. In general term, the probability to move to any state by performing any action is the probability to take some action a and sum all probabilities of states s' reachable from s using a , and finally sum over all such actions

$$Pr(\text{reach any state using any action} | s) = \sum_{a \in A} \pi(a | s) \sum_{s' \in S} Pr(s' | s, a) \quad (2.6)$$

Equivalently, we can write the probability to reach any state using any action and receiving any reward

$$Pr(\text{any state any action any reward} | s) = \sum_{a \in A} \pi(a | s) \sum_{s' \in S} \sum_{r \in R} Pr(s', r | s, a) \quad (2.7)$$

2.4 Goals and Rewards

The agent's goal is to maximize the expected return.

For a finite horizon of size T , the undiscounted sum of rewards is

$$G_t = R_{t+1} + R_{t+2} + \dots + R_{t+T} = \sum_{k=0}^T R_{t+k+1} \quad (2.8)$$

For an infinite horizon, we add a discount factor, as otherwise infinite sum would result in an agent that does not really care for the reward mechanism. As previously stated, the intuition here is reward now $>$ future reward.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (2.9)$$

where $\gamma \in (0, 1)$

2.5 Policies, Value-function and Q-function

The policy π defines our strategy, namely, what we choose to do at every step

$$\pi(s, a) = \Pr[a_t = a | s_t = s] \quad (2.10)$$

The goal of π is to maximize the value function, which is the cumulative expected return of following π starting from some state s

$$V_{\pi}(s) = \mathbb{E}_{\pi}[G_t | s_t = s] = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s \right] \quad (2.11)$$

Notice that the expectation is w.r.t π , meaning that after the initial state $s_t = s$, the next states are fully determined by π . We can think of V_{π} as a measurement of "How good is π ?", as intuitively, we can choose the policy that provides us the maximal expected return.

We can also define the Q -function, which is the same as V except the fact that we start from s and perform an initial action a (that may or may not be one of π 's options)

$$Q_{\pi}(s, a) = \mathbb{E}_{\pi}[G_t | s_t = s, a_t = a] = \mathbb{E}_{\pi} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a \right] \quad (2.12)$$

2.6 The Bellman equation

Theorem 2.6.1 The value function can be written as

$$\begin{aligned} V_{\pi}(s) &= \mathbb{E}_{\pi}[G_t | s_t = s] \\ &= \sum_{a \in A} \pi(a|s) \sum_{s' \in S} \sum_{r \in R} \Pr[s', r | s, a] [r + \gamma V_{\pi}(s')] \\ &= \mathbb{E}_{a \in A} [\mathbb{E}_{s', r} [r + \gamma V_{\pi}(s')]] \end{aligned} \quad (2.13)$$

*Proof.*¹ The reward and time t can be rewritten as

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \quad (2.14)$$

Therefore the can re-write the Value-function as

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} | s_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} | s_t = s] \end{aligned} \quad (2.15)$$

Focusing on the second term, we will use the law of iterated expectation

$$\mathbb{E}[Y | X = x] = \mathbb{E}[\mathbb{E}[Y | X = x, Z = z] | X = x]$$

with $Y = G_{t+1}$, $X = S_t, x = s, Z = S_{t+1}$ and $z = s'$, hence

$$\begin{aligned} \mathbb{E}_\pi[G_{t+1} | s_t = s] &= \mathbb{E}[\mathbb{E}[G_{t+1} | S_t = s, S_{t+1} = s'] | S_t = s] \\ &= \mathbb{E}[\mathbb{E}[G_{t+1} | S_{t+1} = s'] | S_t = s] \\ &= \mathbb{E}[V_\pi(S_{t+1} = s') | S_t = t] \end{aligned} \quad (2.16)$$

In the 2nd transition we removed the inner condition for $S_t = s$ as $G_{t+1} = R_{t+2} + \gamma R_{t+3} + \dots$ does not depend on S_t . This is the case as every reward term R_{t+i} is only a function of the current state and action, so since R_t is not present, no term in G_{t+1} is related to S_t (only to S_{t+1}, S_{t+2}, \dots). In the last transition we use the fact that the inner \mathbb{E} term is nothing but the value function for $t \leftarrow t + 1$.

Substituting all to 2.15 we have

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[R_{t+1} | s_t = s] + \gamma \mathbb{E}_\pi[G_{t+1} | s_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} | s_t = s] + \gamma \mathbb{E}[V_\pi(S_{t+1} = s') | S_t = t] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1} = s') | S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s' \in S} \sum_{r \in R} Pr(s', r | s, a) [r(s, a, s') + \gamma V_\pi(S_{t+1} = s')] \end{aligned} \quad (2.17)$$

R_{t+1} describes the reward obtained when moving from s to s' using a , so it can be written as $r(s, a, s')$. Furthermore, the expectation \mathbb{E}_π is w.r.t to the states, actions and rewards induced by π (so the probability associated with every term is the one introduced in 2.7), and by summing over $r \in R$ we also indicate the fact that the transition $s \rightarrow s'$ could be rewarded with multiple different rewards (more than one option is plausible). ■

The bellman optimality equation is the bellman equation for the optimal Value-function

$$V^*(s) = \max_a \mathbb{E}[r(s, a) + \gamma V^*(s')] \quad (2.18)$$

¹<https://stats.stackexchange.com/questions/243384/deriving-bellmans-equation-in-reinforcement-learning>

Theorem 2.6.2 The Q-function can be written as

$$Q_{\pi}(s, a) = \sum_{s', r} [r(s, a) + \gamma \sum_{a'} Q_{\pi}(s', a')] \quad (2.19)$$

Proof. Not included ■

The bellman optimality equation is the bellman equation for the optimal Q-function

$$Q^*(s, a) = \mathbb{E}[r(s, a) + \gamma \max_{a'} Q^*(s', a)] \quad (2.20)$$

2.7 Policy Iteration

Is a method of finding an optimal policy by performing two steps - evaluation and improvement. After a random initialization of both π and V (Step I), we evaluate the Value-function given some policy π (Step II)². We do this by constantly sampling our environment and updating the value function for every state respectively. The loop stops when the change in value function (for all s) is smaller than some tolerance ϵ . We use the notation $\pi(a|s)$ to indicate the probability $Pr(\pi(s) = a)$.

Next, in step III, we observe the current Value function and choose an action that maximizes it. This will be considered our new returned action for every single state

Finally, we combine policy evaluation and policy improvement in an iterative process. More specifically, we evaluate V and improve π until a fixed point is reached (for all s , $\pi(s)$ has not changed, possibly up to some margin δ)³.

2.8 Value iteration

In some cases it may not be efficient or even possible to scan all states $s \in S$, but scanning all possible actions $a \in A$, is. In VI, we choose the value of some state s as the maximal V function over all $a \in A$ and when the needed tolerance was reached, our returned policy would be the action that maximizes the final V function

2.9 Monte-Carlo

In cases where the dynamics $Pr[s'|s, a]$ and the reward G_t are unknown (model-free setting, 1.2.1), we can use a Monte-Carlo approach to sample the environment and come up with approximations for the value-function and Q-function. To do so, one must make sure that the episodes are finite (the number of transition until termination is $< \infty$). Another important factor is how shall we behave when encountering the same state more than once, and there are two common variants

- First-Visit-Monte-Carlo (FVMC): estimates the return obtained only after the first visit to s (ignore future visits to s)
- Every-Visit-Monte-Carlo (EVMC): estimates the average returns obtained after all visits to s (average all rewards obtained from s in the episode)

²The convergence of PE is the result of the Policy evaluation convergence theory. see [HERE](#) for more info

³I highly recommend checking out [THIS](#) implementation, by Denny Britz

Algorithm 1 Policy Iteration**Require:** tolerance $\varepsilon > 0$ and an MDP $\langle S, A, R, Pr : S \times R \rightarrow \mathbb{R}, \gamma \rangle$ $V, \pi \leftarrow$ Random Initialization $\in \mathbb{R}^{|S|}$

▷ Step I: Initialization

while True **do**

▷ Step II: Policy Evaluation

 $\Delta \leftarrow 0$ **for** $s \in S$ **do** $v \leftarrow V(s)$ $V(s) \leftarrow \sum_{a \in A} \pi(a|s) \sum_{s' \in S} \sum_{r \in R} Pr[s', r|s, a][r + \gamma V(s')]$

▷ using thrm. 2.6.1

 $\Delta \leftarrow \max(\Delta, |v - V(s)|)$ **end for****if** $\Delta < \varepsilon$ **then** break**end if****end while**policy_stable \leftarrow True

▷ Step III: Policy Improvement

for $s \in S$ **do**old_ $\pi \leftarrow \pi(s)$ $\pi(s) \leftarrow \operatorname{argmax}_{a \in A} \sum_{s' \in S} \sum_{r \in R} Pr[s', r|s, a][r + \gamma V(s')]$ **if** old_ $\pi \neq \pi(s)$ **then** policy_stable \leftarrow False**end if****end for****if** policy_stable **then** return V, π **else** go to step II**end if****Algorithm 2** Value Iteration**Require:** tolerance $\varepsilon > 0$ and an MDP $\langle S, A, R, Pr : S \times R \rightarrow \mathbb{R}, \gamma \rangle$ $V \leftarrow$ Random Initialization $\in \mathbb{R}^{|S|}$ **while** True **do** $\Delta \leftarrow 0$ **for** $s \in S$ **do** $v \leftarrow V(s)$ $V(s) \leftarrow \max_{a \in A} \sum_{s' \in S} \sum_{r \in R} Pr[s', r|s, a][r + \gamma V(s')]$ $\Delta \leftarrow \max(\Delta, |v - V(s)|)$ **end for****if** $\Delta < \varepsilon$ **then** break**end if****end while****return** $\pi(s) = \operatorname{argmax}_{a \in A} \sum_{s' \in S} \sum_{r \in R} Pr[s', r|s, a][r + \gamma V(s')]$ for all $s \in S$

2.9.1 Approximating Value-function

To approximate the value function V_π of a given policy π , we will sample a trajectory and update the cumulative discounted reward given the reward we have received. In the case of FVMC, we save the resulted reward for a given state *iff* we did not encounter it previously. The above is also described in alg. 3

Algorithm 3 First-Visit MC for Value function approximation

Require: a policy π to be evaluated

$V(s) \leftarrow$ arbitrary initialization $\in \mathbb{R}$ for all $s \in S$

$Returns(s) \leftarrow$ empty list $[]$ for all $s \in S$

while True **do**

$\tau \leftarrow \{s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T\}$ \triangleright generate an episode using π

$G \leftarrow 0$

for step $t = T - 1, T - 2, \dots, 0$ **do**

$G \leftarrow \gamma G + r_{t+1}$

if $s_t \notin \{s_0, s_1, \dots, s_{t-1}\}$ **then** \triangleright Verifying first-visit

$Returns(s_t).append(G)$

$V(s_t) \leftarrow avg(Returns(s_t))$

end if

end for

end while

2.9.2 Approximating policies

When a model of the world is available, we have already seen (in Policy Iteration, for example) that given the state-values only, one could generate a policy. On the other hand, when the model is not available - the state values alone does not contain enough information to formulate a policy, and one must estimate the action values in order to come up with a good policy. This means that approximating the Q -function could be useful to determine a policy π .

The formalism of FVMC for Q -function approximation is almost identical to alg. 3, except the fact that in a First-Visit, we make sure that both the state s and the action a were not encountered yet in the trajectory. It is also important to understand that if π is deterministic, following it we only observe returns for one specific action from each state. This means that there are no returns to average, hence we in fact do not allow our estimate to explore the state-action space properly. To solve this we must enforce exploration by, for example, setting a random state-action starting point for every episode.

To understand how a policy can be generated, we will go through the following steps:

Consider a Monte-Carlo version of classical policy iteration - Given some initial policy π_0 , we approximate G_π over and over again using MC sampling (with additional starting point exploration). From here, for any action-value function Q , the greedy policy is the one that chooses a maximal action, i.e $\pi(s) = \operatorname{argmax}_a Q(s, a)$. The above is neatly described in the following pseudo-code

2.10 On/Off-Policy methods

There are two types of policy methods

Algorithm 4 MCFV with Exploring Starts (ES) for estimating π_*

```

 $\pi(s) \in A$  arbitrary for all  $s \in S$  ▷ Initialization
 $Q(s, a) \in \mathbb{R}$  arbitrary for all  $s \in S$  and  $a \in A$ 
 $Returns(s, a) \leftarrow$  empty list  $[]$  for all  $s \in S$  and  $a \in A$ 
while True do
   $s_0 \in S$  and  $a_0 \in A$  randomly chosen s.t all pairs  $(s_i, a_i)$  are reachable from  $(s_0, a_0)$  with prob  $> 0$ 
   $\tau \leftarrow \{s_0, a_0, r_1, s_1, a_1, r_2, \dots, s_{T-1}, a_{T-1}, r_T\}$  ▷ generate an episode using  $\pi$  from  $(s_0, a_0)$ 
   $G \leftarrow 0$ 
  for step  $t = T - 1, T - 2, \dots, 0$  do
     $G \leftarrow \gamma G + r_{t+1}$ 
    if  $(s_t, a_t) \notin \{(s_0, a_0), (s_1, a_1), \dots, (s_{t-1}, a_{t-1})\}$  then ▷ Verifying state-action first-visit
       $Returns(s_t, a_t).append(G)$ 
       $Q(s_t, a_t) \leftarrow avg(Returns(s_t, a_t))$ 
       $\pi(s_t) \leftarrow \operatorname{argmax}_a Q(s_t, a)$ 
    end if
  end for
end while

```

- On-policy methods: attempts to evaluate or improve the policy that is being used to make decisions. As stated before, if π does not attain all state-action pairs with probability > 0 , we will poorly explore the space.
- Off-policy methods: attempt to evaluate or improve a policy other than the one used to generate the data (the one that selects actions).

We work with two distinct policies:

- The target policy π - the one that we wish to learn
- the behavior policy b - the one used to generate the data

While On-policy methods tend to be more data efficient, they require new samples with each change of policy. Off-policy, on the other hand, are slower but more powerful and general, as they can be used to learn from various sources (like from a human expert)

2.10.1 Importance sampling for Off-Policy methods

Importance sampling is a technique for estimating expected values under one distribution given samples from another. It is performed by weighting returns according to the fraction of probabilities to a trajectory under some policy.

Lets assume that the behavior policy b is stochastic and the target policy π is deterministic. This means that the trajectories in the data (that were chosen by b) may be different than those chosen by π , which begs the question of how to calculate the expected return? The solution would be to weigh the return based on how it resembled the actual values returned by the target policy.

Consider the trajectory $\tau = \{s_t, a_t, s_{t+1}, a_{t+1}, \dots, s_T\}$. The probability to obtain τ given the starting state s_t and the actions $a_{t:T-1} \sim \pi$ is

$$Pr[\tau | s_t, a_{t:T-1} \sim \pi] = \prod_{k=t}^{T-1} \pi(a_k | s_k) Pr[s_{k+1} | s_k, a_k] \quad (2.21)$$

Denoting the importance sampling for the time window $[t, t+1, \dots, T-1]$ as $\rho_{t:T-1}$, we take the relative probability of the trajectory for the target and behaviour policy

$$\rho_{t:T-1} \equiv \frac{Pr[\tau|s_t, a_{t:T-1} \sim \pi]}{Pr[\tau|s_t, a_{t:T-1} \sim b]} = \prod_{k=t}^{T-1} \frac{\pi(a_k|s_k)}{b(a_k|s_k)} \quad (2.22)$$

Notice that even though the probabilities $P[s'|s, a]$ may be unknown, they cancel out in 2.22. From here, we can use $\rho_{t:T-1}$ and the return G_t of the behaviour policy b to obtain V_π , as

$$V_\pi(s) = \mathbb{E}[\rho_{t:T-1} G_t | s_t = s] \quad (2.23)$$

For example, if some trajectory τ is twice as plausible under b than it is under π , the expected return for π would be $1/2$ (in expectation) the return under b , which can also be seen as $\rho = 1/2$.

From here, we can take the MC algorithm (that averages returns), provide it with episodes following b but still estimate V_π .

Let $\mathcal{T}(s)$ be all time steps state s was visited (over all episodes), and $T(t)$ be the time of termination after time t for a given episode, then $\{G_t\}_{t \in \mathcal{T}(s)}$ is the set of returns associated with s across all episodes, and $\{\rho_{t:T(t)=1}\}_{t \in \mathcal{T}(s)}$ are the corresponding IS ratios. To estimate V_π we can use

$$V_\pi(s) = \frac{\sum_{t \in \mathcal{T}(s)-1} \rho_{t:T(t)=1} G_t}{|\mathcal{T}(s)|} \quad (2.24)$$

2.11 ϵ -Greedy Algorithms

In RL we often need to balance between

- Exploration - experimenting with multiple actions to better assess the expected reward
- Exploitation - attempt to maximize the reward by choosing the optimal action. We can do this by, for example, estimating our Q function with MC as

$$\hat{Q}_t(a, s) = \frac{1}{N_t(a)} \sum_{i=1}^T (r_i | a_t = a, s_t = s)$$

where $N_t(a)$ is the number of times the action a was chosen, and then choose an action

$$a_t^* = \operatorname{argmax}_{a \in A} \hat{Q}_t(a, s)$$

Notice how these two may collide, as when we explore the environment we also may not always choose an optimal action. In an ϵ -greedy approach, we explore with probability ϵ , and in all other cases we choose the optimal action:

- With probability $1 - \epsilon$: select $a_t^* = \operatorname{argmax}_{a \in A} \hat{Q}_t(a, s)$
- with probability ϵ : choose a random $a \in A$

Intuitively speaking, as we maintain the ability to explore forever, we will eventually find an optimal policy. We also make sure to not include the randomness in the testing phase.

2.12 Temporal Difference (TD) Learning

When discussing MC learning, we showed how sampling the environment without knowing the dynamics of the system could be enough to learn V , Q , and a policy π directly. On the other hand,

when using MC we must make sure that the episodes are finite, and we could only learn based on complete episodes. In TD learning, on the other hand, both infinite environments and incomplete sequences learning is possible, as we can update our approximation after every step (compared to after every episode in MC). We define the TD-error as the difference between the optimal value function V_t^* and the current prediction V_t :


$$\begin{aligned}
 \delta_t^{TD}(s) &= V_t^*(s) - V_t(s) \\
 &= \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} - V_t(s) \\
 &= r_{t+1} + \sum_{k=1}^{\infty} \gamma^k r_{t+k+2} - V_t(s) \\
 &= r_{t+1} + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+k+2} - V_t(s) \\
 &= r_{t+1} + \gamma V_{t+1}^* - V_t(s)
 \end{aligned} \tag{2.25}$$

As we do not know V_{t+1}^* , we can approximate it using the predicted V_{t+1}

$$\delta_t^{TD} \approx r_{t+1} + \gamma V_{t+1} - V_t \tag{2.26}$$

This error is used as the general update rule, where we also add a learning rate α (much like in gradient descent, where we can think of $V_{t+1} - V_t$ as the term ∇V_t)

$$V_t \leftarrow V(t) + \alpha [r_{t+1} + \gamma V_{t+1} - V_t] \tag{2.27}$$

 notice that we do not use an expectation term (as seen in policy iteration for example), as the update rule is the result of looking only one step into the future, given some episode rollout.

Algorithm 5 One-Step TD [TD(0)]

Require: π the policy to evaluate, $\alpha \in \mathbb{R}$

$V(s) \in \mathbb{R}$ arbitrary initialized for all $s \in S$

for each episode E **do**

▷ sampling a trajectory like in MC

for each $s \in E$ **do**

$a \leftarrow \pi(s)$

$r, s' \leftarrow$ take action a and observe r, s'

▷ taking one step to the future

$V(s) \leftarrow V(s) + \alpha(r + \gamma V(s') - V(s))$

▷ using eq. 2.27

$s \leftarrow s'$

end for

end for

return V


2.12.1 On-Policy TD Control: SARSA

SARSA is an on-policy (remember 2.10) algorithm that used TD(0) in order to approximate the Q -function. The idea will be similar to the value function update error, and the MC sampling would

Algorithm 6 SARSA

Require: $\alpha \in \mathbb{R}$
 $Q(s, a) \in \mathbb{R}$ arbitrary initialized for all $s \in S$ and for all $a \in A$
for each episode E **do** ▷ sampling a trajectory like in MC
 for each $s \in E$ **do**
 choose a from s given Q ▷ like in ϵ -greedy (2.11)
 $r, s' \leftarrow$ take action a and observe r, s' ▷ taking one step to the future
 choose a' from s' given Q ▷ like in ϵ -greedy (2.11)
 $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma Q(s', a') - Q(s, a))$
 $s \leftarrow s', a \leftarrow a'$
 end for
end for
return Q

consider both the next state *and* the next action. Do notice that, as always, to approximate Q we must find its optimal value for every $s \in S$, and $a \in A$ - which is computationally difficult.

 the name SARSA stems from the idea that the update rule uses the quintuple $\langle s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1} \rangle$

2.12.2 Off-Policy TD Control: Q-Learning

In Q-Learning, we do not look at the next $Q(s', a')$, but on the Q -function that has maximal value over all possible actions $\max_a Q(s', a)$. This means that the corresponding maximal Q -function was provided given an action that may or may not be the result of our policy π (hence, it is Off-Policy). More specifically - Q-learning is based on a greedy approximation of the optimal policy - which is the behaviour policy (compared to SARSA, that only used the current policy). Notice that we still use the current policy, as it determines which state-action pairs are visited and updated.

Algorithm 7 Q-Learning

Require: $\alpha \in \mathbb{R}$
 $Q(s, a) \in \mathbb{R}$ arbitrary initialized for all $s \in S$ and for all $a \in A$
for each episode E **do** ▷ sampling a trajectory like in MC
 for each $s \in E$ **do**
 choose a from s given Q ▷ like in ϵ -greedy (2.11)
 $r, s' \leftarrow$ take action a and observe r, s' ▷ taking one step to the future
 $Q(s, a) \leftarrow Q(s, a) + \alpha(r + \gamma \max_a Q(s', a) - Q(s, a))$
 $s \leftarrow s'$
 end for
end for
return Q

Also notice we did not track down the next action a' , as we did not use it (scanned all $a \in A$ instead). It should also be stated that Q-Learning usually converges quicker, due to the optimal choice of $\max_a Q(s', a)$. Having said that, as we do not take into consideration the next state a' , using ϵ -greedy actions might mean that we take a step into a state with very bad reward (like falling down a cliff), simply as we are less conservative (due to exploration and not admitting to a next action).

3. DQN & it's derivatives

This section is based on sources I found online (because the lecture was not uploaded). Our new goal would be to use a parameterized function approximator to represent the state-action Q-function, instead of representing it with a table. In other words, we wish to find a function $\hat{Q}(s, a, \theta) \approx Q(s, a)$, where θ is our function's parameters and $Q(s, a)$ is the true function/oracle.

Lets start of with the ideal assumption, in which the oracle $Q(s, a)$ is accessible. Our function approximator could be learned using SGD, that is, by minimizing the squared loss $J(\theta)$ w.r.t to the oracle over batches sampled from our environment

$$J(\theta) = \mathbb{E} [(Q(s, a) - \hat{Q}(s, a, \theta))^2] \quad (3.1)$$

We update our parameter vector θ using $\theta \leftarrow \theta - \Delta\theta$, where

$$\begin{aligned} \Delta\theta &= -\frac{1}{2}\alpha \nabla_{\theta} J(\theta) \\ &= -\frac{1}{2}\alpha \nabla_{\theta} \mathbb{E} [(Q(s, a) - \hat{Q}(s, a, \theta))^2] \\ &= -\frac{1}{2}\alpha \cdot 2\mathbb{E} [Q(s, a) - \hat{Q}(s, a, \theta)] \cdot (-\nabla_{\theta} \hat{Q}(s, a, \theta)) \\ &= \alpha \mathbb{E} [Q(s, a) - \hat{Q}(s, a, \theta)] \nabla_{\theta} \hat{Q}(s, a, \theta) \end{aligned} \quad (3.2)$$

As $Q(s, a)$ is generally unknown, it must be replaced with some approximated target. Recall that in SARSA for example, our target was based on the temporal difference $r + \gamma \hat{Q}(s', a', \theta)$. In classic Q-learning, on the other hand, we used an off policy approach, in which our target was $r + \gamma \max_{a' \in A} \hat{Q}(s', a', \theta)$. As we now proceed to describe DQN, we will follow the update rule of

$$\Delta\theta = \alpha \mathbb{E} \left[r + \gamma \max_{a' \in A} \hat{Q}(s', a', \theta) - \hat{Q}(s, a, \theta) \right] \nabla_{\theta} \hat{Q}(s, a, \theta) \quad (3.3)$$

3.1 Deep Q-Network

We now ask how shall we describe our approximated $\hat{Q}(s, a, \theta)$, and the answer is a deep neural network.

3.1.1 Architecture

In the original paper, the DQN architecture was based on convolutional neural network. The network takes in an input of shape $84 \times 84 \times 4$ (a processed batch of images, which is considered the state)¹, and propagates this input via three convolutional layers. To finally come up with an action value, there are two fully connected layer, where the last one has a single output for every possible action. Notice how our function approximates an action for every state, hence can be written as $\forall s \in S : \hat{Q}(s, \theta) : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|A|}$.

3.1.2 Training DQN

Those of us who are familiar with supervised learning, might think that the kind of "true-label" introduced in 3.3 is not much of a true label at all, as it is based on the network itself, which is constantly changing. This fact makes the learning process unstable, therefore few "tricks" were introduced in the paper - the *Experience Replay*, and a separate *Target Network*. Summarizing the two in a short sentence, our Q-network is learned by minimizing

$$J(\theta) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1}} \left[y_t^{DQN} - \hat{Q}(s_t, a_t, \theta) \right] \quad (3.4)$$

where y_t^{DQN} is the one-step TD component

$$y_t^{DQN} = r_t + \gamma \max_{a' \in A} \hat{Q}(s_{t+1}, a', \theta^-) \quad (3.5)$$

Notice how we use both θ and θ^- , where θ is the original parameters vector of the network but θ^- represents the parameters of the target network. By doing so we update θ based on values from a previous version of our model. Furthermore, the batches we sample are provided from all past transition tuples.

Experience Replay

We store the transitions $\langle s_t, a_t, r_t, s_{t+1} \rangle := e_t$ at each time-step in a fixed buffer $D_t = \{e_1, e_2, \dots, e_t\}$, and During SGD, we only sample uniformly from D . This is because

- Data efficiency: each experience e_i can potentially be used in many updates
- De-correlation: randomly sampling leads to de-correlation between consecutive experiences. As correlation breaks, expected values will fluctuate less, meaning that the variance in sampling is reduced, hence stability is increased.
- Smoother divergence: As our training process is based on a large number of experiences, outliers tend to average out with the rest of the samples, leading to less oscillations in the training process.

Note that in a more sophisticated experience replay, we might weigh experiences based on their importance and keep the relevant ones for longer.

¹Originally, images were $210 \times 160 \times 3$, but the preprocessing applied spacial dimension reduction to 84×84 , extracted the Y channel from the RGB and concatenated 4 consecutive images as "memory"

Target Network

Another stability improvement comes from the fact that we use two neural nets. Ideally, we would like to minimize the effect of our targets being *Non-Stationary*, and we do so by setting a network that is only updated after $C \gg 1$ steps. This means that our non-stationary target will, in a sense, be stationary for at least C steps, hence stabilizing our learning process even further.

Algorithm 8 Deep Q-Learning (DQN)

```

Initialize experience replay  $D$  with fixed size
Initialize network  $\hat{Q}$  with random weights  $\theta$ 
Initialize target network  $\hat{Q}_T$  with random weights  $\theta^- = \theta$ 
for episode  $m = 1, 2, \dots, M$  do
    observe environment to get  $s_1$ 
    take action  $a_t \leftarrow \begin{cases} \text{random,} & \text{with probability } \varepsilon \\ \operatorname{argmax}_{a \in A} \hat{Q}(s_t, a, \theta), & \text{otherwise} \end{cases}$ 
     $r_t, s_{t+1}, done_i \leftarrow$  Take action  $a_t$  in the environment
     $D \leftarrow D \cup \{(s_t, a_t, r_t, s_{t+1}, done_i)\}$ 
     $B \leftarrow$  sample a random batch  $\{(s_i, a_i, r_i, s_{i+1}, done_i)\}_{i=1}^N$  from  $D$ 
    for every experience  $\langle s_i, a_i, r_i, s_{i+1}, done_i \rangle \in B$  do
         $y_i \leftarrow r_i$  if  $done_i$  else  $r_i + \gamma \max_{a' \in A} \hat{Q}_T(s_{i+1}, a', \theta^-)$ 
    end for
    perform SGD on  $J(\theta) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{Q}(s_i, a_i, \theta))^2$  w.r.t  $\theta$ 
    every  $C$  steps  $\theta^- \leftarrow \theta$ 
end for

```

3.2 Double Deep Q-Network (DDQN)

We'd like to state that the max operator is prone to over estimation. Consider the following example:²

■ **Example 3.1** Say $N \gg 1$ people have equal weight of 80kg , and we'd like to measure all of their weights using a weighing scale that is off by ± 1 kg (equal probability to measure > 80 and < 80). Lets run two sets of measurements:

- Denote the weight measurement of the i 'th person as X_i , and set $Y = \max_i X_i$. We can intuitively understand that almost surely $Y > 80$, as almost surely there exists some j for which $X_j > 80$. This really tells us that the max operator is prone to over estimation when noise is introduced to the system.
- As a second experiment, we will measure each person's weight twice and store the values in X_1^i, X_2^i . to estimate Y , we first calculate $n = \operatorname{argmax}_i X_1^i$. Next, we take the second measurement $Y = X_2^n$ as our maximal value. Notice that as X_2^n is independent from X_1^n , it is equally likely to overestimate the real value or underestimate it, hence it is not systematically over-optimistic. So everyone weighs 80kg and $X_1^n > 80$ with high probability, but X_2^n is both > 80 and < 80 with even probability.

■

²from CIS 522 YouTube channel

With the example above in mind, notice how in the DQN algorithm we both choose an action a_t using $\hat{Q}(s_t, a, \theta)$ and evaluate it (when calculating y_t), which means that we tend to overestimating the target values. To address this issue we replace the current update of y_i (when not *done* _{i}) with

$$y_i^{DoubleDQN} = r_i + \gamma \hat{Q}\left(s_{i+1}, \underset{a' \in A}{\operatorname{argmax}} \hat{Q}(s_{i+1}, a', \theta), \theta^-\right) \quad (3.6)$$

In other words, we choose an action (argmax) using a network with parameters θ (that is currently being trained), but evaluate the action using a network with parameters θ^- (that is not being trained)

3.3 Dueling network

Starting off with a new definition, the *Advantage* function that is associated with a policy π and a state-action function Q is given by

$$A^\pi(s, a) = Q^\pi(s, a) - V_\pi(s) \quad (3.7)$$

We can think of the advantage as some measurement to how important is some action, as in - take the importance of a given state-action pair ($Q(s, a)$) and subtract the importance of the state ($V(\pi)$) to get the importance of a given action. Note that as $V^\pi(s) = \mathbb{E}_{a \sim \pi}[Q^\pi(s, a)]$ (see eq. 1.6), we have that $\mathbb{E}_{a \sim \pi}[A^\pi(s, a)] = 0$.

A *Dueling* network approximates the Q -function by first separating the output of the network into two distinct elements - one is $V(s)$ and the other is $A(s, a_1), A(s, a_2), \dots$, and then extracts from both the value of Q . Recall that in regular DQN, for every state we estimated the value of all action

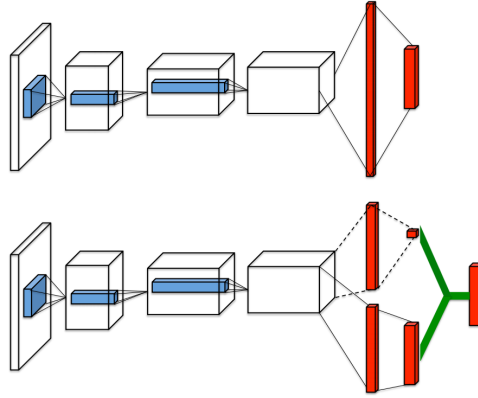


Figure 3.1: (Top) a standard DQN architecture, where the last layer is a vector representing $Q(s, a_1), Q(s, a_2), \dots$ (Bottom) a Dueling DQN, where the upper branch represents the value of $V(s)$, the lower branch represents the values of $A(s, a_1), A(s, a_2), \dots$ and the final layer represents $Q(s, a_1), Q(s, a_2), \dots$. As the input/output are the same, Alg. 8 can be applied given the Dueling architecture as well.

choices. This may be unnecessary in cases where an action has little to no effect on the outcome. For example - if you fell down a cliff, it is really irrelevant what are the action values of steering the wheel. As we split our architecture to two distinct components, we can learn state values without having to learn the effect of each action on those states - as those may be an unnecessary computation for some states.


3.3.1 Implementation

Denoting the output of the upper branch as $V(s, \theta, \theta_V)$ and the output of the lower branch as $A(s, \theta, \theta_A)$ where θ are shared parameters and θ_V, θ_A are distinct parameters for every branch, it may seem reasonable to add these values to obtain $Q(s, a, \theta, \theta_V, \theta_A)$. The immediate problem that arises is that given such Q , we cannot recover the exact values of both A and V , as for example it may be the case that $Q = (A + x) + (V - x)$, and x can be chosen freely.

To distinguish A and V , the following formula was applied:

$$\hat{Q}(s, a, \theta, \theta_A, \theta_V) = V(s, \theta, \theta_V) + \left[A(s, a, \theta, \theta_A) - \max_{a' \in \text{Actions}} A(s, a', \theta, \theta_A) \right] \quad (3.8)$$

This trick forces the Q value associated with the maximizing action to equal V (as for the maximizing action, which is the action that is choose, the squared brackets zero out). This means that the upper stream (V) can be identified as the Q function value, and the lower stream would be the advantage function.

 Alternatively, one can use

$$\hat{Q}(s, a, \theta, \theta_A, \theta_V) = V(s, \theta, \theta_V) + \left[A(s, a, \theta, \theta_A) - \frac{1}{|\text{Actions}|} \sum_{a' \in \text{Actions}} A(s, a', \theta, \theta_A) \right] \quad (3.9)$$

which was shown to improve stability, as was stated in the article



4. Policy Gradients

(see also 1.3).

Instead of value function of state-action value function approximation, we can aim to model the policy itself. Again, we take the path of function approximation, rather than a look-up table approach that assign every state an action. Generally speaking, we define a parameterized policy function π_θ and try to optimize it using a gradient ascent optimization process

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t) \quad (4.1)$$

where $\nabla J(\theta_t)$ is an estimated performance measurement.

π_θ itself may have various kinds of outputs, given the problem setup (see 1.1.3 for more info)

- For a discrete action space - $\pi_\theta(s) \in \mathbb{R}^{|A|}$, where each entry may represent a probability associated with an action
- For continuous action space - the output may be mean and covariance of a Gaussian distribution (from which we sample an action), or a floating point value as commonly used in regression tasks

Policy-gradient methods can learn an appropriate balance between exploration/exploitation, hence are preferred over ε -greedy approaches. This also means that they provide us with a smoother action selection, as changes in final action choice are based in changes to the parameters of the model, rather than some random value exceeding ε

4.1 The Policy-Gradient theorem

Let us define the performance measure $J(\theta)$ as the value function associated with the parameterized policy

$$J(\theta) = V_{\pi_\theta}(s) \quad (4.2)$$

The PG theorem states that the gradient of $J(\theta)$ is proportional to the gradient of our policy, weighted by the Q function over all states, and by the average number of times each state is visited

in a trajectory $\mu(s)$.

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a Q_{\pi_\theta}(s, a) \nabla \pi_\theta(a|s) \quad (4.3)$$

The proof can be seen in S&B2020, page 325.

The PG theorem provides us with an analytic term to the change in performance w.r.t the parameters of our policy.

4.2 The REINFORCE Algorithm

If we follow a target policy π , the weighted sum over all states in eq. 4.3 can be replaced with expected value over the policy, as the number of times $\mu(s)$ some state was reached under π is the same as summing all values under the policy associated with the state s . In other words, $\nabla J(\theta) = \mathbb{E}_\pi [\sum_a Q_{\pi_\theta}(s, a) \nabla \pi_\theta(a|s)]$.

Supposedly, this is enough to formulate a gradient-ascent algorithm, as the mean value is replaced with sampling a set of experiences - $\theta_{t+1} = \theta_t + \alpha \sum_a \hat{Q}(s_t, a, \theta_Q) \nabla \pi_\theta(a|s_t)$. Having said that, this approach means that an approximation of Q is needed as well, and the update rule takes into consideration *all* actions, which is not ideal.

A preferred option would be to articulate a learning scheme that is based only on taking one step in our environment, i.e taking an action a_t and understanding from it alone how to change the weights:

$$\begin{aligned} \nabla J(\theta) &= \mathbb{E}_\pi \left[\sum_a Q_{\pi_\theta}(s_t, a) \nabla \pi_\theta(a|s_t) \right] \\ &= \mathbb{E}_\pi \left[\sum_a \pi_\theta(a|s_t) Q_{\pi_\theta}(s_t, a) \frac{\nabla \pi_\theta(a|s_t)}{\pi_\theta(a|s_t)} \right] \text{ [multiply and divide the same value]} \\ &= \mathbb{E}_\pi \left[Q_{\pi_\theta}(s_t, a_t) \frac{\nabla \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \right] \text{ [replace } \sum_a [\dots] \text{ with sampling } a_t \sim \pi] \\ &= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)} \right] [Q_\pi(s_t, a_t) = \mathbb{E}_\pi[G_t|s_t, a_t]] \end{aligned} \quad (4.4)$$

Hence we can now rewrite the update rule as $\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}$. Notice that this approach is only applicable in the episodic setting, as to calculate the return G_t we need to sum all future rewards until the end is reached. Intuitively, we can decompose this expression as

- $\nabla \pi_\theta(a_t|s_t)$ is a vector pointing to the direction (in parameter space) that increases the probability of taking a_t the most, given the state s_t
- the "amplitude" is $\frac{G_t}{\pi_\theta(a_t|s_t)}$, meaning that the update tends to increase G_t - actions are chosen if they maximize the return, and decrease $\pi_\theta(a_t|s_t)$ - this makes sense as otherwise, maximizing $\pi_\theta(a_t|s_t)$ means that actions that are frequent (high $\pi_\theta(a_t|s_t)$) will be chosen even if they do not yield high return.

One last trick is to notice that $\nabla \ln x = \nabla x/x$, so the final update rule is

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla \ln \pi_\theta(a_t|s_t) \quad (4.5)$$

From here, we can generate an episode from our initialized policy function, calculate the return as a sum of discounted (or undiscounted) reward (remember that we assume finite episodes setting) and update the network parameters using the update rule from above.

Algorithm 9 REINFORCE: MC Policy-Gradient control (episodic case) for control**Require:** a (differentiable), randomly initialized policy $\pi_\theta(a|s)$ **Require:** learning rate α **while** did not converge **do** generate an episode $\{s_0, a_0, r_1, \dots, s_{T-1}, a_{T-1}, r_T\}$ given π_θ **for** each step $t = 0, 1, \dots, T - 1$ **do** $G \leftarrow \sum_{k=t+1}^T \gamma^{k-t-1} r_k$ $\theta \leftarrow \theta + \alpha \gamma^t G \nabla \ln \pi_\theta(a_t|s_t)$ **end for****end while****4.2.1 REINFORCE with Baseline**

Sampling an episode may cause fluctuations and slow convergence properties. To mitigate the problem we can include a baseline, which can be any function that we compare our predicted return to. We use this baseline in the update rule as $\theta_{t+1} = \theta_t + \alpha(G_t - b(s_t)) \frac{\nabla \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}$. We also must ensure that our baseline does not depend on the actions, as those are what we aim to learn. More specifically, a reasonable baseline would be the value function $\hat{V}(s_t, \theta_V)$ (recall that V is the mean value of G , starting from s_t and following the policy π). \hat{V} should be calculated using some method that was previously discussed.

Algorithm 10 REINFORCE: MC Policy-Gradient control (episodic case) for control**Require:** a (differentiable), randomly initialized policy $\pi_\theta(a|s)$ **Require:** a (differentiable), randomly initialized value function $\hat{V}(s, \theta_V)$ **Require:** learning rate α_V, α **while** did not converge **do** generate an episode $\{s_0, a_0, r_1, \dots, s_{T-1}, a_{T-1}, r_T\}$ given π_θ **for** each step $t = 0, 1, \dots, T - 1$ **do** $G \leftarrow \sum_{k=t+1}^T \gamma^{k-t-1} r_k$ $\delta \leftarrow G - \hat{V}(s_t, \theta_V)$ $\theta_V \leftarrow \theta_V + \alpha_V \delta \nabla \hat{V}(s_t, \theta_V)$ $\theta \leftarrow \theta + \alpha \gamma^t \delta \nabla \ln \pi_\theta(a_t|s_t)$ **end for****end while**

Intuitively, $G_t - \hat{V}(s_t)$ tells us how much we did after time step t (G_t) compared to what was expected to achieve ($\hat{V}(s_t)$). If we did better than expected, the log likelihood of that given action is now proportional to "how much better" we actually performed. If we anticipated the return exactly, then the actions we chose are ideal, hence no more updates are required ($G(t) - b(s_t) = 0$). Comparing this to the previous method, where we always multiplied by the actual return G_t , unnecessary updates would have happened, which hinders convergence.