Throughout we are working with a background probability space denoted $(\Omega, \mathcal{H}, \mathbb{P})$.

0.1 Reinforcement Learning

Through the probability space, denote this

In Reinforcement Learning (RL) we are concerned with finding an optimal policy for an agent in some environment. Typically (also in the case of Q-learning) this environment is a Markov decision process

Definition 1. A Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ consists of

- $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a measurable space of **states**.
- $(\mathcal{A}, \Sigma_{\mathcal{A}})$ a measurable space of **actions**.
- $P(\cdot \mid \cdot) : \Sigma_{\mathcal{S}} \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$ a probability kernel (of **transition** probabilities).
- $R(\cdot \mid \cdot) : \mathbb{B} \times (\mathcal{S} \times \mathcal{A}) \to [0, 1]$ a probability kernel (of **reward** probabilities).
- $\gamma \in (0, 1)$ a **discount** factor.

In order for this to make sense we here include

Definition 2 (Probability kernel)**.** Let $(X, \Sigma_X), (Y, \Sigma_Y)$ be measurable spaces. A function $\kappa(\cdot \mid \cdot) : \Sigma_Y \times X \to [0, 1]$ is a **probability kernel** provided

- $B \mapsto \kappa(B \mid x) \in \mathcal{P}(\Sigma_Y)$ that is $\kappa(\cdot \mid x)$ is a probability measure for any $x \in X$.
- $x \mapsto \kappa(B \mid x) \in \mathcal{M}(\Sigma_X, \Sigma_Y)$ that is $\kappa(B \mid \cdot)$ is $(\Sigma_X/\Sigma_Y$-) measurable for any $B \in \Sigma_Y$.

Note that both P and R to be stochastic and that R can depend on the action as well as the state. This is perhaps the most general way to define an MDP, generalizing some definitions. Common variations include that R depends on $\mathcal{S}$ only, R is deterministic, or P is deterministic.

Definition 3 (Policy)**.** A (**randomized**, **stationary**) **policy** π is probability kernel

$$\pi(\cdot \mid \cdot) : \Sigma_{\mathcal{A}} \times \mathcal{S} \to [0, 1]$$

An MDP together with a policy and an initial distribution $\mu \in \mathcal{P}(\mathcal{S})$ give rise to a countable stochastic process, $(X_i)_{i \in \mathbb{N}} = (S_i, A_i, R_i)_{i \in \mathbb{N}}$ that is a probability measure P^{π}_{μ}. on $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^{\mathbb{N}}$. See [ref. to Feinberg On Meas. and Repr. of Str. Meas. p. 31-32, and pos. Ionescu Tulcea] for full details of how this is constructed. Intuitively S_1 is drawn from μ, then for all $i \in \mathbb{N}$ A_i is drawn from $\pi(\cdot \mid S_i)$, a reward is then drawn from $R(\cdot \mid S_i, A_i)$, then S_{i+1} is drawn from $P(\cdot \mid S_i, A_i)$ and so on. We let E^{π}_{μ} denote the expectation taken w.r.t P^{π}_{μ}. When μ is a Dirac measure δ_x, i.e. $\mu(\{x\}) = 1$ for some x, we shall generally write x instead of μ, E.g. $\mathbb{E}^{\pi}_{s}$ the expectation taken with respect to $P^{\pi}_{\delta_s}$

0.2 Q-Learning

Fix an MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and a policy π. We assume from now on that any $R \sim R(\cdot \mid s, a)$ is bounded with $|R| \leqslant R_{\max}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

ne has the reduced the problem of finding a good strategy to choosing the action that leads to the highest value. Of course this value of a state will depend strongly on the policy being followed.

The **ideal** value function w.r.t. π, $V^{\pi} : \mathcal{S} \to \mathbb{R}$

$$V^{\pi}(s) := \mathbb{E}^{\pi}_{s} \sum_{t=1}^{\infty} \gamma^t R_t$$

where R_t are the projections of the random process generated from the MDP with starting distribution δ_s onto the rewards of each step. This is well-define because each R_t is bounded by $R_{\max}$ so $V^{\pi}(s) \leqslant \sum_{t=0}^{\infty} \gamma^t R_{\max} = \frac{1}{1-\gamma} R_{\max} := V_{\max} < \infty$. The value function gives the expected accumulated reward when starting in state s and following policy π.

The **ideal** Q-function w.r.t. π, $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$Q^\pi(s,a) := r(s,a) + \gamma \mathbb{E}(V^\pi(S) \mid S \sim P(\cdot \mid s,a))$$

where $r(s,a) := \mathbb{E}(R \mid R \sim R(\cdot \mid s,a))$.

One could think that it is a bit superfluous to define both a value and an action value function. According to [todo: ref] the main reason to work with Q-functions is that it is more difficult to work with value function for several reasons. Firstly to know what is the best action a^* given a state s and a value function V, one has to calculate for each action a the distribution of the next state s' and take expectation over $V(s')$. This either requires full knowledge of the transition kernel (this falls outside the so called model-free approaches) or some way of estimating it, and in both cases at computational cost. Whereas Q-functions simply requires finding $\operatorname{argmax}_{a \in \mathcal{A}} Q(s,a)$.

We define the operator P^π

$$(P^\pi Q)(s,a) := \mathbb{E}(Q(S',A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot \mid S'))$$

Intuitively this operator yields the expected action value function when looking *one step ahead* following the policy π and taking expectation of Q. Note that $\|P^\pi Q\|_\infty \leqslant \|Q\|_\infty$. $\|\cdot\|_\infty$.

We define the operator T^π called the Bellman operator by

$$(T^\pi Q)(s,a) := r(s,a) + \gamma (P^\pi Q)(s,a)$$

The Bellman operator adjusts an action value function Q to look more like Q^π. This is intuitively by making one iteration of reward-propagation discounting with γ. And indeed

Proposition 1. Q^π is the unique fixed point of T^π.

Proof. Q^π is a fixed point because for any $s \in \mathcal{S}, a \in \mathcal{A}$

$$\begin{aligned}
T^\pi Q^\pi(s,a) &= r(s,a) + \gamma (P^\pi Q^\pi)(s,a) \\
&= r(s,a) + \gamma \mathbb{E}(Q^\pi(S',A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot, S')) \\
&= r(s,a) + \gamma \mathbb{E}(r(S',A') + \gamma \mathbb{E}(V^\pi(S'') \mid S'' \sim (\cdot \mid S',A')) \mid S' \sim P(\cdot \mid s,a), A' \sim \pi(\cdot, S')) \\
&= r(s,a) + \gamma \mathbb{E}(V^\pi(S') \mid S' \sim P(\cdot \mid s,a)) \\
&= Q^\pi(s,a)
\end{aligned}$$

Now since

$$Q^\pi - T^\pi Q = T^\pi Q^\pi - T^\pi Q = \gamma P^\pi (Q^\pi - Q)$$

by induction

$$Q^\pi - (T^\pi)^n Q = (\gamma P^\pi)^n (Q^\pi - Q)$$

And since γP^π contracts to 0, in fact every bounded Q-function converges to Q^π when iteratively applying T^π. In particular Q^π is the only fixed point of T^π. □

If an action value function Q satisfies that $Q(s,\mathcal{A})$ has a greatest value for every $s \in \mathcal{S}$ then we can define **greedy** policy π with respect to Q to be a policy choosing an action with maximal value of Q for each state. That is $\pi(s) = \delta_a$ for some $a \in \operatorname{argmax}_a Q(s,a)$. We then write $\pi = \pi_Q$.

Let π_0 be a policy and $Q_0 = Q^{\pi_0}$ be its ideal Q-function. One can now consider the greedy policy $\pi_1 = \pi_{Q_0}$. Note that

$$\begin{aligned}
T^{\pi_1} Q^{\pi_0} &= r(s,a) + \gamma \mathbb{E}\left(\sup_{a' \in \mathcal{A}} Q^{\pi_0}(S',a') \mid S' \sim P(\cdot \mid s,a) \right) && (1) \\
&\geqslant r(s,a) + \gamma \mathbb{E}\left(Q^{\pi_0}(S',A') \mid S' \sim P(\cdot \mid s,a), A' \sim \pi_0(\cdot \mid S') \right) && (2) \\
&= Q^{\pi_0} && (3)
\end{aligned}$$

so applying T^{π_1} iteratively on Q_0 creates a monotonically increasing sequence of Q-functions which by **??** converge to Q^{π_1}, proving that $Q^{\pi_1} \geqslant Q^{\pi_0}$. One can then repeat the process with π_1 and obtain a increasing sequences of (ideal) Q-functions with associated policies $(Q_0, Q_1, \dots), (\pi_0, \pi_1, \dots)$. Variations of this idea is called **policy iteration** and has been studied a lot. Variations include

stopping the "value iteration" (applying the T^{π_i} operator) at various stages before again updating the policy to be greedy w.r.t. to the next Q-function. An important special case is where we simply alternate between updating the policy and the Q-function in every step. We can capture this in a single operator called the Bellman *optimality* operator T, defined as

$$TQ := T^{\pi_Q} Q$$

The optimal Q-function is defined as

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

where the supremum is taken over all policies.

Note that V^π, Q^π and Q^* are usually infeasible to calculate to machine precision, unless $\mathcal{S} \times \mathcal{A}$ is finite and not very big.

0.3 Artificial Neural Networks

Definition 4. An **ANN** (Artificial Neural Network) with structure $\{d_i\}_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij} : \mathbb{R} \to \mathbb{R})_{j=1}^{d_i}$ and weights $\{W_i \in M^{d_i \times d_{i-1}}, v_i \in \mathbb{R}^{d_i}\}_{i=1}^{L+1}$ is the function $F : \mathbb{R}^{d_0} \to \mathbb{R}^{d_{L+1}}$

$$F = w_{L+1} \circ \sigma_L \circ w_L \circ \sigma_{L-1} \circ \cdots \circ w_1$$

where w_i is the affine function $x \mapsto W_i x + v_i$ for all i.

Here $\sigma_i(x_1, \ldots, x_{d_i}) = (\sigma_{i1}(x_1), \ldots, \sigma_{id_i}(x_{d_i}))$.

$L \in \mathbb{N}_0$ is called the number of hidden layers.

d_i is the number of neurons or nodes in layer i.

An ANN is called *deep* if there are two or more hidden layers.

Definition 5 (Sparse ReLU Networks)**.** For $s, V \in \mathbb{R}$ a (s,V)-**Sparse ReLU Network** is an ANN f with any structure $\{d_i\}_{i\in[L+1]}$, all activation functions being *ReLU* i.e. $\sigma_{ij} = \max(\cdot, 0)$ and any weights (W_ℓ, v_ℓ) satisfying

- $\max_{\ell\in[L+1]} \left\| \widetilde{W}_\ell \right\|_\infty \leqslant 1$
- $\sum_{\ell=1}^{L+1} \left\| \widetilde{W}_\ell \right\|_0 \leqslant s$
- $\max_{j\in[d_{L+1}]} \|f_j\|_\infty \leqslant V$

Here $\widetilde{W}_\ell = (W_\ell, v_\ell)$. The set of them we denote $\mathcal{F}\left(L, \{d_i\}_{i=0}^{L+1}, s, V\right)$.

0.4 Fitted Q-Iteration

We here present the algorithm which everything in this paper revolves around:

Algorithm 1: Fitted Q-Iteration Algorithm

Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, function class $\mathcal{F}$, sampling distribution ν, number of iterations K, number of samples n, initial estimator $\widetilde{Q}_0$

for $k = 0, 1, 2, \ldots, K-1$ **do**

Sample i.i.d. observations $\{(S_i, A_i), i \in [n]\}$ from ν obtain $R_i \sim R(S_i, A_i)$ and $S_i' \sim P(S_i, A_i)$

Let $Y_i = R_i + \gamma \cdot \max_{a\in\mathcal{A}} \widetilde{Q}_k(S_i', a)$

Update action-value function:

$$\widetilde{Q}_{k+1} \leftarrow \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(S_i, A_i))^2$$

Define π_K as the greedy policy w.r.t. $\widetilde{Q}_K$

Output: An estimator $\widetilde{Q}_K$ of Q^* and policy π_K