In this section we will develop general theory about decision processes and value function (including Q-functions) that is used across all sources considered in this paper, including the question of optimal policy existence.

0.0.1 History dependent decision process

We define in this section a quite general framework. We do this partly in the quest to have a united framework to talk about results from a variety of sources, and relate them to each other in generality. And partly to avoid defining various concepts such as value functions everytime a new context is considered. A source which uses a setup which is almost as general can be found in [ref. to Schal]. In this section recall that $\underline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\}$, $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ and $\underline{\overline{\mathbb{R}}} = \mathbb{R} \cup \{\pm\infty\}$.

Definition 0.1 (History dependent decision process)**.** A **history dependent decision process** (HDP) is determined by

1. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})_{n\in\mathbb{N}}$ a measurable space of **states** for each timestep.

2. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})_{n\in\mathbb{N}}$ a measurable space of **actions** for each timestep.

for each $n \in \mathbb{N}$ we define the so called **history** spaces

$$\mathcal{H}_1 = \mathcal{S}_1, \quad \mathcal{H}_2 = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2, \quad \mathcal{H}_3 = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \underline{\overline{\mathbb{R}}} \times \mathcal{A}_2 \times \mathcal{S}_3$$

$$\mathcal{H}_n = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \underline{\overline{\mathbb{R}}} \times \mathcal{A}_2 \times \mathcal{S}_3 \times \underline{\overline{\mathbb{R}}} \times \cdots \times \mathcal{S}_n$$

$$\mathcal{H}_\infty = \mathcal{S}_1 \times \mathcal{A}_1 \times \mathcal{S}_2 \times \underline{\overline{\mathbb{R}}} \times \dots$$

with associated product σ-algebras

3. $(P_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{H}_n \times \mathcal{A}_n \rightsquigarrow \mathcal{S}_{n+1}$ kernels called the **transition** kernels.

4. $(R_n)_{n\in\mathbb{N}}$ a sequence of $\mathcal{H}_{n+1} \rightsquigarrow \underline{\overline{\mathbb{R}}}$ kernels called the **reward** kernels.

The name *decision process* is used for many different processes across litterature but many of them generalize to the above. Some authors also use the name *dynamic progamming model.*

Notice the slight irregularity in the beginning of the history spaces: We are missing a reward state after $\mathcal{S}_1$. This could have been avoided by introducing some start reward but we will do without.

Assumption 1. (Reward independence) P_n, R_n and policies are only allowed to depend on the past states and actions, and not the rewards.

In all sources known to this writer assumption 1 is assumed. This is a bit of a puzzle since it is obvious that one could want to define algorithms (policies) that take into account which rewards they received in the past. We will also do this but stick to the standard and never attempt to evaluate ideal value functions of policies that depend on rewards. Thus we will let assumption 1 hold from now on and throughout this paper.

The majority of sources considered in this paper also specialize with the following:

Assumption 2 (One state and action space)**.** $\mathcal{S}_1 = \mathcal{S}_2 = \ldots := \mathcal{S}$, $\mathcal{A}_1 = \mathcal{A}_2 = \ldots := \mathcal{A}$

We will do without this for the rest of this section in order to present some results in the generality they deserve. Later we will look at settings which do not specialize this way. One could ask if it is possible to embed the general decision process into one with assumption 2 by setting $\mathcal{S} := \bigcup_{i\in\mathbb{N}} \mathcal{S}_i$ and $\mathcal{A} := \bigcup_{i\in\mathbb{N}} \mathcal{A}_i$ or similar. One attempt at this can be found in [**?**] chapter 10, but this will not be covered here.

Other ways to specialize include reducing one or both of the transition and reward kernels to functions defined on $\mathcal{S} \times \mathcal{A}$. These processes are often called *deterministic*, but the exact definitions vary across sources, and we will instead specify each setting individually.

For a decision process we can define

Definition 0.2 (Policy)**.** A (randomized) **policy** $\pi = (\pi_n)_{n\in\mathbb{N}}$ is a sequence of $\mathcal{H}_n \rightsquigarrow \mathcal{A}_n$ kernels. The set of all policies we denote $R\Pi$. The policy π is called **semi Markov** if each π_i only depends on the first and last state in the history and is called **Markov** if only the last. The sets are denoted $sM\Pi$ and $M\Pi$. Furthermore π is called **deterministic** if all π_i are degenerate, i.e. for all i we have $\pi_i(\{a_i\} \mid h_i) = 1$ for some $a_i \in \mathcal{A}_i$. Under assumption 2 it makes sense to make a (Markov) policy $(\pi, \pi, \dots)$, where π only depends on the last state. Such a policy is called **stationary**, and the set of them denoted $S\Pi$. We denote the deterministic version of the policy classes by the letter D.

We have the following inclusions

$$\begin{array}{cccc} S\Pi & \subseteq M\Pi & \subseteq sM\Pi & \subseteq R\Pi \\ \rotatebox{90}{$\subseteq$} & \rotatebox{90}{$\subseteq$} & \rotatebox{90}{$\subseteq$} & \rotatebox{90}{$\subseteq$} \\ DS\Pi & \subseteq DM\Pi & \subseteq DsM\Pi & \subseteq D\Pi \end{array}$$

Proposition 0.3. A dynamic progamming model together with a policy π defines a probability kernel $\kappa_\pi : \mathcal{S}_1 \to \mathcal{H}_\infty$.

Proof. This is the Ionescu-Tulcea kernel generated by $\dots R_2 P_2 \pi_2 R_1 P_1 \pi_1$. □

This kernel yields a probability measure $\kappa_\pi \mu$ on $\mathcal{H}_\infty$ for every $\mu \in \mathcal{S}_1$. In particular for any $s \in \mathcal{S}_1$ $\kappa_\pi \delta_s$ yields the measure $\kappa_\pi(\cdot \mid s)$ and we shall occasionally write this $\kappa_\pi s$ and integration with respect to it $\mathbb{E}_s^\pi$.

Across litterature generally any function mapping a state space $\mathcal{S}$ to $\overline{\mathbb{R}}$ can be called a (state) **value** function. Similarly any $\overline{\mathbb{R}}$ valued function on pairs of states and actions can be called (state) **action value** or **Q-** function. The idea behind such functions are commonly to estimate the cumulative rewards associated with a state or state-action pair and the trajectory of states it can lead to. In order to be able to sum rewards we will need one of the following conditions:

Condition F^- (Reward finity from above)**.** $\int_{[0,\infty]} x \,\mathrm{d}R_i(x \mid h) < \infty$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

Condition F^+ (Reward finity from below)**.** $\int_{[-\infty,0]} x \,\mathrm{d}R_i(x \mid h) > -\infty$ for all $h \in \mathcal{H}_{i+1}$ and $i \in \mathbb{N}$

The letter F comes from [**?**]. When assuming either of (F^+) or (F^-) we ensure that the summation of finitely many rewards has a well defined mean in $\overline{\mathbb{R}}$, and then the following definition makes sense

Definition 0.4 (Finite horizon value function)**.** Let $\underline{R}_i : \mathcal{H}_\infty \to \overline{\mathbb{R}}$ be the projection onto the ith reward. Define

$$V_{n,\pi}(s) = \mathbb{E}_s^\pi \sum_{i=1}^{n} \underline{R}_i$$

called the kth finite horizon value function. When $n = 0$ we say $V_{0,\pi} = V_0 := 0$ for any π.

The finite horizon value function measures the expected total reward of starting in state s and then follow the policy π for n steps. This way it measures the *value* of that particular state given a policy and *horizon* (number of steps). We would like to extend this to an infinite horizon value function, i.e. letting n tend to ∞. To ensure that the integral is well-defined we need one of the following conditions

Condition P (Reward non-negativity)**.** $R_i([0,\infty] \mid h) = 1, \forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Condition N (Reward non-positivity)**.** $R_i([-\infty, 0] \mid h) = 1 \; \forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Condition D (Discounting)**.** There exist a bound $R_{\max} > 0$ and a $\gamma \in [0,1)$ called the **discount** factor such that $R_i([-R_{\max}\gamma^i, R_{\max}\gamma^i]) = 1 \; \forall h \in \mathcal{H}_{i+1}, i \in \mathbb{N}$

Again the letters P, N and D are adopted from [**?**].

Definition 0.5. We define the infinite horizon value function by

$$V_\pi(s) = \mathbb{E}_s^\pi \lim_{n\to\infty} \sum_{i=1}^n \underline{R}_i$$

The infinite horizon value function V_π measures the expected total reward after following the policy π an infinite number of steps.

Remark 0.6. Whenever we are working with the finite horizon value function we will always assume that either (F^+) or (F^-) holds without stating this explicitly. If a result only holds under e.g. (F^+) we will of course be explicit about this be marking it accordingly with a (F^+).

Similarly whenever we work with the infinite horizon value function we will always assume that at least one of (P), (N) or (D) holds. We will mark propositions and theorems by e.g. (D) (P) when the result only holds for if discounting *or* reward non-negativity is assumed. Note that obviously (P) implies F^+ and (N) implies F^-.

Remark 0.7. Since we are under assumption 1, when talking about the finite or infinite value functions, we can actually reduce the reward kernels to functions $r_i : \mathcal{H}_{i+1} \to \overline{\mathbb{R}} = h \mapsto \int r \; \mathrm{d}R_i(r \mid h)$ (note that r_i is measurable due to **??**). Another way of stating this is that the value functions are indifferent to whether we use deterministic or stochastic rewards. This however does not mean that we can dispose completely of stochastic rewards, as they still make a difference to model-free algorithms that do not know the reward kernel, and therefore cannot simply integrate it.

For use later we mention some properties of these value functions.

Proposition 0.8. When well-defined the value functions $V_{n,\pi}, V_\pi$ are measurable into $(\overline{\mathbb{R}}, \overline{\mathbb{B}})$.

Proof. Use **??**. □

Proposition 0.9. $\lim_{n\to\infty} V_{n,\pi} = V_\pi$ for all $\pi \in R\Pi$.

Proof. By monotone or dominated convergence. □

Proposition 0.10. Under (D) for any $\pi \in R\Pi$ we have
$\left|V_{n,\pi}\right|, \left|V_\pi\right| \leqslant R_{\max}(1-\gamma) < \infty$.

Proof. For any $\pi \in R\Pi$

$$|V_\pi(s)| \leqslant \mathbb{E}_s^\pi \sum_{i\in\mathbb{N}} |\underline{R}_i| \leqslant \sum_{i\in\mathbb{N}} \gamma^{i-1} R_{\max} = R_{\max}/(1-\gamma)$$

This also covers $V_{n,\pi}$. □

As this bound will occur again and again we denote it

$$V_{\max} := R_{\max}(1-\gamma)$$

Optimal policies

Let $(\mathcal{S}_n, \mathcal{A}_n, P_n, R_n)_{n\in\mathbb{N}}$ be a decision process.

Definition 0.11 (Optimal value functions)**.**

$$V_n^*(s) := \sup_{\pi \in R\Pi} V_{n,\pi}(s) \qquad\qquad V^*(s) := \sup_{\pi \in R\Pi} V_\pi(s)$$

This is called the **optimal value function** (and the nth optimal value function). A policy $\pi^* \in R\Pi$ for which $V_{\pi*} = V^*$ is called an **optimal policy**. If $V_{n,\pi*} = V_n^*$ it is called n-optimal.

Proposition 0.12. (D)
$\left|V_k^*\right|, |V^*| \leqslant V_{\max}$.

Proof. By proposition 0.10 all terms in the suprema are within this bound. □

Remark 0.13. It is known that the optimal value function might not be Borel measurable (see ex. 2 p. 233 [**?**]). Perhaps this is not suprising since we are taking a supremum over sets of policies which have cardinality of at least the continuum. However it is often possible to show that they are. We will take these discussions as they occur in various settings.

At this point some central questions can be asked.

1. To which extend does an optimal policy π^* exist?
2. Does V_n^* converge to V^*?
3. When can optimal policies be chosen to be Markov, deterministic, etc.?
4. Can an algorithm be designed to efficiently find V^* and π^*?

These questions has been answered in a variety of settings. We will try to address them in order by strength of assumptions they require.

Schäls theorem

In a quite general setting, questions 1 and 2 is investigated in [**?** , **?** (**?**)]. Here some additional structure on our process is imposed:

Setting 1 (Schäl)**.** 1. $V_\pi < \infty$ for all policies $\pi \in R\Pi$.

2. $(\mathcal{S}_n, \Sigma_{\mathcal{S}_n})$ is assumed to be standard Borel. I.e. $\mathcal{S}_n$ is a non-empty Borel subset of a Polish space and $\Sigma_{\mathcal{S}_n}$ is the Borel subsets of S_n.
3. $(\mathcal{A}_n, \Sigma_{\mathcal{A}_n})$ is similarly assumed to be standard Borel.
4. $\mathcal{A}_n$ is compact.
5. $\forall s \in \mathcal{S}_1 : Z_n = \sup_{N \geqslant n} \sup_{\pi \in R\Pi} \sum_{t=n+1}^{N} \mathbb{E}_s^\pi r_n \to 0$ as $n \to \infty$.

In this setting Schäl introduced two set of criteria for the existence of an optimal policy:

Condition S. 1. The function

$$(a_1, a_2, \dots, a_n) \mapsto P_n(\cdot \mid s_1, a_1, s_2, a_2, \dots, s_n, a_n)$$

is set-wise continuous (hence the name **S**) for all $s_1, \dots, s_n \in \mathcal{S}^{\underline{n}}$.

2. r_n is upper semi-continuous.

Condition W. 1. The function

$$(h_n, a_n) \mapsto P_n(\cdot \mid h_n, a_n)$$

is weakly continuous (hence the name **W**).

2. r_n is continuous.

Theorem 0.14 (Schäl)**.** When either (S) or (W) hold then

1. There exist an optimal policy $\pi^* \in R\Pi$.
2. $V_n^* \to V^*$ as $n \to \infty$.

Proof. We refer to [**?**]. □

Schäls theorem tells us that optimal policies exist in a wide class of decision processes. However in many cases we are looking at processes in which the next state in independent of the history. In such cases it makes sense to ask if optimal policies can be chosen within the system of policy subclasses. Such questions will be addressed in the next section.

0.0.2 The Markov decision process and its operators

Definition 0.15 (Markov decision process)**.** A **Markov decision process** (MDP) consists of

1. $(\mathcal{S}, \Sigma_{\mathcal{S}})$ a measurable space of states.
2. $(\mathcal{A}, \Sigma_{\mathcal{A}})$ a measurable space of actions.
3. $P : \mathcal{S} \times \mathcal{A} \rightsquigarrow \mathcal{S}$ a transition kernel.
4. $R : \mathcal{S} \times \mathcal{A} \rightsquigarrow \overline{\mathbb{R}}$ a reward kernel.
5. An optional disount factor $\gamma \in [0, 1]$ (when not discounting put $\gamma = 1$).

This is a special case of the history dependent decision process (definition 0.1) with

- Assumption 2 is satisfied i.e. $\mathcal{S}_1 = \mathcal{S}_2 = \cdots = \mathcal{S}, \quad \mathcal{A}_1 = \mathcal{A}_2 = \cdots = \mathcal{A}$.
- P_n depends only on s_n and a_n and does not differ with n, i.e. $P_n(\cdot \mid s_1, \dots, s_n, a_n) = P(\cdot \mid s_n, a_n)$ for all $n \in \mathbb{N}$.
- R_n depends only on s_n and a_n and does not differ with n except for a potential discount. I.e. $R = R_n/\gamma^{n-1}$ for all $n \in \mathbb{N}$

We will write P instead of P_n understanding kernel compositions as if using P_n.

At this point it makes sense to define

Definition 0.16 (The T-operators)**.** For a stationary policy π and measurable $V : \mathcal{S} \to \overline{\mathbb{R}}$ with $V \geqslant 0$, $V \leqslant 0$ or $|V| < \infty$ we define the operators

$$P_\pi V := s \mapsto \int V(s') \,\mathrm{d}P\pi(s' \mid s)$$

$$T_\pi V := s \mapsto \int r(s,a) + \gamma V(s') \,\mathrm{d}(P\pi)(a, s' \mid s)$$

$$TV := s \mapsto \sup_{a \in \mathcal{A}} T_a V(s)$$

where $T_a = T_{\delta_a}$.

Proposition 0.17 (Properties of the T-operators)**.** Let $\pi = (\pi_1, \pi_2, \dots)$ be a Markov policy.

1. The operators P_π, T_π and T commutes with limits.
2. $V_{k,\pi} = T_{\pi_1} V_{k-1,(\pi_2,\dots)} = T_{\pi_1} \dots T_{\pi_k} V_0$.
3. $V_\pi = \lim_{k\to\infty} T_{\pi_1} \dots T_{\pi_k} V_0$
4. If π is stationary $T_\pi V_\pi = V_\pi$.
5. (D) T and T_π are γ-contractive on $\mathcal{L}_\infty(\mathcal{S})$.
6. (D) V_π is the unique bounded fixed point of T_π in $\mathcal{L}_\infty(\mathcal{S})$

Proof.

1. By monotone or dominated convergence theorems.
2. $$\begin{aligned}
&T_{\pi_1} V_{k,(\pi_2,\dots)}(s_1)\\
&= \int r(s_1,a_1) + \gamma \int \sum_{i=2}^{k+1} \gamma^{i-2} r(s_i,a_i) \,\mathrm{d}\kappa_{\pi_2,\dots}(a_2, s_3, a_3, \dots \mid s_2) \,\mathrm{d}P\pi_1(a_1, s_2 \mid s_1)\\
&= \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \,\mathrm{d} \dots P\pi_2 P\pi_1(a_1, s_2, \dots \mid s_1)\\
&= \int \sum_{i=1}^{k+1} \gamma^{i-1} r(s_i,a_i) \,\mathrm{d}\kappa_\pi(a_1, s_2, \dots \mid s_1)\\
&= V_{k+1,\pi}(s_1)
\end{aligned}$$

   Now use this inductively.
3. This is by 2. and proposition 0.9.
4. By 3. $T_\pi V_\pi = T_\pi \lim_{k\to\infty} T_\pi^k V_0 = \lim_{k\to\infty} T_\pi^{k+1} V_0 = V_\pi$.
5. Let $V, V' \in \mathcal{L}_\infty(\mathcal{S})$ and let $K = \left\| V - V' \right\|_\infty$. Then since the rewards are bounded

   $$\left| T^\pi V - T^\pi V' \right| = \gamma \left| \int V(s') - V'(s') \,\mathrm{d}P\pi(s' \mid s) \right| \leqslant \gamma K$$

   For T use the same argument and the fact that $\left| \sup_x f(x) - \sup_y g(y) \right| \leqslant |\sup_x f(x) - g(x)|$ for any $f, g : X \to \mathbb{R}$.
6. By 4., 5. and Banach fixed point theorem.

□

0.0.3 Q-functions

Definition 0.18. Let $\pi \in R\Pi$. Define

$$Q_{k,\pi}(s,a) = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_{k,\pi}, \qquad Q_\pi = r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V_\pi$$

$$Q_k^* = \sup_{\pi \in R\Pi} Q_{k,\pi}, \qquad Q^* = \sup_{\pi \in R\Pi} Q_\pi$$

Define $Q_0 = r$ then we make the convention that $Q_0^* = Q_{0,\pi} = Q_0 = r$.

The idea of Q-functions (and the letter Q) originates to [**?** , **?** (**?**)]. Upon the definition he notes

> "This is much simpler to calculate than $[V_\pi]$ for to calculate $[Q_\pi]$ it is only necessary to look one step ahead [...]"

A clear advantage of working with Q-function $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ rather than a value function $V : \mathcal{S} \to \overline{\mathbb{R}}$, is that finding the optimal action in state s requires only a maximization over the Q-function itself: $a = \operatorname{argmax}_{a \in \mathcal{A}} Q(s,a)$. This should be compared to finding a best action according to a value function V: $a = \operatorname{argmax}_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{P(\cdot|s,a)} V$. Besides being less simple, this requires taking an expectation with respect to both the reward and transition kernel. Later we will study settings where we are not allowed to know the process kernels when attempting to find the optimal strategy. In these situations the advantage of Q-functions is clear. For now however the transition kernel will remain known and we will in this section see how the results of state-value functions translate to Q-functions. The results in this section are original in the generality here presented, as I was unable to find them elsewhere.

Proposition 0.19. (D)
$\lim_{k \to \infty} Q_{k,\pi} = Q_\pi$. Furthermore it holds that $|Q_{k,\pi}|, |Q_\pi|, |Q_k^*|, |Q^*| \leqslant V_{\max}$.

Proof. By dominated convergence or monotone convergence and proposition 0.10. □

In parallel to the operators for state-value functions we define

Definition 0.20 (T operators for Q-functions)**.** For any stationary policy $\pi \in S\Pi$ and measurable $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ with $Q \geqslant 0, Q \leqslant 0$ or $|Q| < \infty$ we define

$$P_\pi Q(s,a) = \int Q(s',a') \, \mathrm{d}\pi P(s',a' \mid s,a)$$

$$T_\pi Q = r + \gamma P_\pi Q$$

$$TQ(s,a) = r(s,a) + \gamma \int \sup_{a' \in \mathcal{A}} Q(s',a') \, \mathrm{d}P(\cdot \mid s,a)$$

where $T_a = T_{\delta_a}$.

Proposition 0.21 (Properties of T-operators for Q-functions)**.** Let $\pi = (\pi_1, \pi_2, \dots) \in M\Pi$ be a Markov policy and $\tau \in S\Pi$ stationary.

1. $T_\tau Q_{k,\pi} = r + \gamma \mathbb{E} T_\tau V_{k,\pi}$
2. $Q_{k,\pi} = T_{\pi_1} \dots T_{\pi_k} Q_0$.
3. $T_\tau Q_\tau = Q_\tau$.
4. (D) T_τ is γ-contractive on $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ and Q_τ is the unique fixed point of T_τ in $\mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$.

Proof.

1. This is essentially due to properties of the kernels. The idea is sketched here
$$T_\mu Q_{k,\pi} = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P \, \mathrm{d}\mu P = r + \gamma \int r + \gamma V_{k,\pi} \, \mathrm{d}P\mu \, \mathrm{d}P = r + \gamma \int T_\mu V_{k,\pi} \, \mathrm{d}P$$
2. Use 1. iteratively starting with $\mu = \pi_1, \pi = (\pi_2, \pi_3, \dots)$.
3. By 2. $T_\pi Q_\pi = T_\pi(r+\gamma\mathbb{E}\lim_{k\to\infty} T_\pi^k V_0) = \lim_{k\to\infty} T_\pi(r+\gamma\mathbb{E}T_\pi^k V_0) = \lim_{k\to\infty}(r+\gamma\mathbb{E}T_\pi^{k+1}V_0) = r + \gamma\mathbb{E}\lim_{k\to\infty} T_\pi^{k+1}V_0 = r + \gamma\mathbb{E}V_\pi = Q_\pi$.
4. The contrativeness of T_π follows from the same argument as for value functions. 2. and Banach fixed point theorem does the rest.

□

Definition 0.22. Let $\pi : \mathcal{S} \rightsquigarrow \mathcal{A}$ be a stationary policy. Define $A_s = \operatorname{argmax}_{a\in\mathcal{A}} Q(s,a)$. If there exist a measurable subset $B_s \subseteq A_s$ for every $s \in \mathcal{S}$ such that
$$\pi\left(B_s \mid s\right) = 1$$
then π is said to be **greedy** with respect to Q and is denoted π_Q.

Proposition 0.23. For any integrable $Q : \mathcal{S} \times \mathcal{A} \to \overline{\mathbb{R}}$ if π_Q is greedy with respect to Q then $T_{\pi_Q} Q = TQ$.

Proof.
$$\begin{aligned} T_{\pi_Q} Q &= r + \gamma \int Q(s,a) \, \mathrm{d}\pi P(s,a \mid \cdot) \\ &= r + \gamma \int\int Q(s,a) \, \mathrm{d}\pi_Q(a \mid s) \, \mathrm{d}P(s \mid \cdot) \\ &= r + \gamma \int \max_{a\in\mathcal{A}} Q(s,a) \, \mathrm{d}P(s \mid \cdot) \\ &= TQ \end{aligned}$$

□

0.0.4 Bertsekas-Shreve framework

The theory described here is largely based on the text book *Stochastic Optimal Control: Discrete-time Case* by [**?** , **?** (**?**)]. Their framework is cost-based as opposed to the this paper reward-based outset. This means that (P) and (N), upper and lower semicontinuity, suprema and infima, ect. are opposite to the source.

Setting 2 (BS)**.**

1. We consider an MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ (see definition 0.15).
2. $\mathcal{S}$ and $\mathcal{A}$ are Borel spaces.
3. $\mathcal{A}$ is compact.
4. $P(S \mid \cdot)$ is continuous for any $S \in \Sigma_{\mathcal{S}}$.

5. $r(s,a) = \gamma^{1-i} \int x \, \mathrm{d}R(x \mid s,a)$ is upper semicontinuous and uniformly bounded from above (least upper bound denoted $0 < R_{\max} < \infty$).

6. The policies must consist of universally measurable probability kernels.

The original setup in [**?**] is slightly different than the setup here presented. Besides having a state and action space, it also features a non-empty Borel space called the *disturbance space* W, a *disturbance kernel* $p : \mathcal{S} \times \mathcal{A} \to W$, instead of a transition kernel which on the other hand is a deterministic *system function* $f : \mathcal{S} \times \mathcal{A} \times W \to \mathcal{S}$ which should be Borel measurable. Moreover it allows for constrains on the action space for each state. This is made precise by a function $U : \mathcal{S} \to \Sigma_{\mathcal{A}}$ and a restriction on $R\Pi$ that all policies π should satisfy $\pi(U(s) \mid s) = 1$. Lastly the rewards are interpreted as negative costs, and thus g is required to be semi *lower*continuous.

By setting $P(\cdot \mid s,a) = f(s,a,p(\cdot \mid s,a))$ and maximizing rewards of upper semicontinuous instead of minimizing lower semicontinuous ones, we fully capture all aspects of the original process and its results, except the for the action constrains.

Notice that setting 2 implies (F^+). Throughout this section are always assumed.

Proposition 0.24. Let $\mathcal{X}, \mathcal{Y}$ be separable and metrizable, $\kappa : \mathcal{X} \to \mathcal{Y}$ be a continuous probability kernel and $f : \mathcal{X} \times \mathcal{Y} \to \overline{\mathbb{R}}$ be Borel-measurable satisfying one of $f \leqslant 0, f \geqslant 0, |f| < \infty$. If f is bounded from above (below) and upper (lower) semicontinuous then

$$x \mapsto \int f \, \mathrm{d}\kappa(\cdot \mid x)$$

is bounded from above (below) and upper (lower) semicontinuous.

Proof. We refer to [**?**] prop. 7.31. □

Proposition 0.25 (Prop. 8.6 in BS)**.** $V_k^* = T^k V_0$ and is upper semicontinuous. Furthermore there exists a sequence of deterministic, stationary, Borel-measurable policies $\tau_1^*, \tau_2^*, \dots \in DS\Pi$ such that $\pi_k^* = (\tau_k^*, \dots, \tau_1^*)$ is k-optimal.

Theorem 0.26 (Cor. 9.17.2 in BS)**.** Under (N) or (D) $V^* = \lim_{k\to\infty} V_k^*$ and is upper semicontinuous. Furthermore there exist a deterministic stationary, Borel-measurable policy π^*.

Analytic setting

For comparison, we include here an similar result in an alternative setting, also considered by [**?**].

Setting 3 (BS Analytic)**.** The same as setting 2 except: P is not necessarily continuous. r is upper semianalytic. $\mathcal{A}$ is not necessarily compact, but there exists a $k \in \mathbb{N}$ such that $\forall \lambda \in \mathbb{R}, n \geqslant k, s \in \mathcal{S}$

$$A_n^\lambda(s) = \left\{ a \in \mathcal{A} \;\middle|\; r(s,a) + \gamma \int V_n^* P(\cdot \mid s,a) \geqslant \lambda \right\}$$

is a compact subset of $\mathcal{A}$.

This setting 3, was actually more widely discussed in [**?**]. We have put more emphasis on the semicontinuous setting, as it appears restrictive to assume the semianalytical property.

Theorem 0.27 (Prop. 9.17 BS)**.** Under setting 3 we have $V^* = \lim_{n\to\infty} V_n^*$ for all $s \in \mathcal{S}$ and there exists a optimal policy π^* which is stationary and deterministic.

Proof. We refer to [**?** , **?** (**?**)] prop. 9.17. □

Implications for value-functions

Let setting 2 hold.

Proposition 0.28. $V^* = V_{\pi^*} = T_{\pi^*}V^* = TV^*$

(D) V^* is the unique fixed point of T in $\mathcal{L}_\infty(\mathcal{S})$.

Proof. Since π^* is optimal $V^* = V_{\pi^*}$ which by proposition 0.17 equals $T_{\pi^*}V_{\pi^*}$. By theorem 0.26 and proposition 0.25 $TV^* = T\lim_{k\to\infty} T^kV_0 = \lim_{k\to\infty} T^{k+1}V_0 = V^*$. If (D) holds $V^* \in \mathcal{L}_\infty(\mathcal{S})$ so by proposition 0.17 5. and 6. we are done. □

Proposition 0.29.

1. $Q_k^* = r + \gamma\mathbb{E}V_k^*$ and is upper semicontinuous.
2. (N) (D) $Q^* = r + \gamma\mathbb{E}V^*$ and is upper semicontinuous.
3. (N) (D) $\sup_{a\in\mathcal{A}} Q^*(s,a) = V^*(s)$.
4. (N) (D) $Q^* = \lim_{k\to\infty} Q_k^*$.
5. (N) (D) $Q^* = Q_{\pi^*}$.

Proof.

1. Since V_k^* is measurable due to proposition 0.25 we see that $Q_k^* = \sup_{\pi\in R\Pi}(r + \gamma\mathbb{E}V_{k,\pi}) \leqslant r + \gamma\mathbb{E}V_k^* = r + \gamma\mathbb{E}V_{\pi_k^*} \leqslant Q_k^*$. Proposition 0.24 gives upper semicontinuity.
2. Since V^* is measurable due to theorem 0.26. Now follow the argument for 1.
3. Let $s \in \mathcal{S}$ then $\sup_{a\in\mathcal{A}} Q^*(s,a) = \sup_{a\in\mathcal{A}}(r(s,a) + \gamma\mathbb{E}_{P(\cdot|s,a)}V^*) = TV^*(s) = V^*(s)$.
4. By monotone or dominated convergence and theorem 0.26.
5. By proposition 0.28 and 2. $Q^* = r + \gamma\mathbb{E}V^* = r + \gamma\mathbb{E}V_{\pi^*} = Q_{\pi^*}$.

□

Proposition 0.30.

1. $TQ_k^* = r + \gamma\mathbb{E}TV_k^*$ and if $\pi^* = (\pi_1^*, \pi_2^* \dots)$ is k-optimal then $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r = T^k r$.
2. $TQ^* = r + \gamma\mathbb{E}TV^*$ and $TQ^* = Q^*$.
3. (D) T is γ-contractive on $\mathcal{L}_\infty(\mathcal{S}\times\mathcal{A})$ and Q^* is the unique fixed point of T in $\mathcal{L}_\infty(\mathcal{S}\times\mathcal{A})$.

Proof.

1.

$$\begin{aligned}
TQ_k^*(s,a) &= T(r + \gamma\mathbb{E}V_k^*)(s,a) \\
&= r(s,a) + \gamma \int \sup_{a'\in\mathcal{A}} (r(s',a') + \gamma\mathbb{E}_{P(\cdot|s',a')}V_k^*) \, \mathrm{d}P(s' \mid s,a) \\
&= r(s,a) + \gamma \int \sup_{a'\in\mathcal{A}} \left(r(s',a') + \gamma \int V_k^*(s'') \, \mathrm{d}P(s'' \mid s',a') \right) \, \mathrm{d}P(s' \mid s,a) \\
&= r(s,a) + \gamma \int TV_k^*(s') \, \mathrm{d}P(s' \mid s,a)
\end{aligned}$$

To get $Q_k^* = T^k r$ use this inductively $Q_k^* = r + \gamma\mathbb{E}V_k^* = r + \gamma TV_{k-1}^* = TQ_{k-1}^* = \dots$. The statement $Q_k^* = T_{\pi_1^*} \dots T_{\pi_k^*} r$ is from proposition 0.21.

2. The argument from 1. also implies this first statement in 2. Now $TQ^* = r + \gamma \mathbb{E} TV^* = r + \gamma \mathbb{E} V^* = Q^*$ by proposition 0.28.

3. The argument is similar to proposition 0.17 pt. 5.

□

Corollary 0.31. (D)

For any $Q \in \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$ $T^k Q$ converges to Q^* with rate γ^k. That is

$$\left\| T^k Q - Q^* \right\|_\infty \leqslant \gamma^k \|Q - Q^*\|_\infty$$

Proof. This is directly from proposition 0.30 pt. 3. □

Proposition 0.32.

1. Let π_i be greedy w.r.t. Q^*_{i-1} then $(\pi_i, \pi_{i-1}, \dots, \pi_1)$ is i-optimal for any $i \in \mathbb{N}$.

2. (N) (D) Any greedy strategy for Q^* is optimal and such exist.

Proof. 1. Such greedy policies exist because $Q_{k,\pi}$ is upper semicontinuous by proposition 0.25. For induction base observe that $Q_{1,\pi_1} = T_{\pi_1} Q_0 = TQ_0 = Q^*_1$. Now assume $Q_{i-1,\pi_{i-1},\dots,\pi_1} = Q^*_{i-1}$. Then $Q_{i,(\pi_i,\dots,\pi_1)} = T_{\pi_i} Q_{i-1,(\pi_{i-1},\dots,\pi_1)} = T_{\pi_i} Q^*_{i-1} = TQ^*_{i-1} = Q^*_i$.

2. Since Q is upper semicontinuous in the second entry the set $A_s = \operatorname{argmax}_{a \in \mathcal{A}} Q(s,a)$ is nonempty and measurable for all s. Pick (by axiom of choice) an $a_s \in A_s$ for every $s \in \mathcal{S}$. Then $\pi(\cdot \mid s) = \delta_{a_s}$ is greedy with respect to Q.

□

Remark 0.33. Most of the results of this section hold also under setting 3 with the addition that 'semicontinuous' is replaced by 'semianalytic'.

0.0.5 Theoretical Q-iteration

Based on the results established so far we can as a non-practical example design the following algorithm:

Algorithm 1: Simple theoretical Q-iteration

```
Input: MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
$\forall (s,a) \in \mathcal{S} \times \mathcal{A} : r(s,a) \leftarrow \int x \, \mathrm{d}R(x \mid s,a)$.
$\widetilde{Q}_0 \leftarrow r$
for $k = 0, 1, 2, \dots, K-1$ do
    $\forall (s,a) \in \mathcal{S} \times \mathcal{A} : \widetilde{Q}_{k+1}(s,a) \leftarrow r(s,a) + \gamma \int \sup_{a' \in \mathcal{A}} \widetilde{Q}_k(s',a') \, \mathrm{d}P(s' \mid s,a)$
Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

Proposition 0.34. (D)

The output $\widetilde{Q}_K$ of algorithm 1 converges to the optimal Q-function Q^* with rate γ^K concretely $\left\| \widetilde{Q}_K - Q^* \right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. This is by corollary 0.31. □

Finite Q-iteration

We have shown how if one knows the dynamics of a stationary decision process satisfying rather broad criteria, such as continuity and compactness, the optimal policy and state-value function can be found simply by iteration over the T-operator and picking a greedy strategy (see proposition 0.34). Of course this is practical computationally, only if the resulting Q functions can be represented and computed in finite space and time. An obvious situation in which such a representation and computation is possible, is the finite case.

Assumption 3. $\mathcal{S} \times \mathcal{A}$ is finite.

Say $|\mathcal{S}| = k$ and $|\mathcal{A}| = \ell$. In this case the transition operator P can be represented as a matrix of *transition probabilities*

$$P := \begin{pmatrix} P(s_1 \mid s_1, a_1) & \dots & P(s_k \mid s_1, a_1) \\ \vdots & \vdots & \vdots \\ P(s_1 \mid s_k, a_\ell) & \dots & P(s_k \mid s_k, a_\ell) \end{pmatrix}$$

then the algorithm becomes

Algorithm 2: Simple finite Q-iteration

```
Input: MPD $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, number of iterations $K$
Set $r \leftarrow \left(\int r \, \mathrm{d}R(r \mid s_1, a_1), \dots, \int r \, \mathrm{d}R(r \mid s_k, a_\ell)\right)^T$
and $\widetilde{Q}_0 \leftarrow r$.
for $k = 0, 1, 2, \dots, K-1$ do
    Set $m(\widetilde{Q}_k) \leftarrow (\max_{a \in \mathcal{A}} Q(s_1, a), \dots, \max_{a \in \mathcal{A}} Q(s_k, a))^T$
    Update action-value function:
        $\widetilde{Q}_{k+1} \leftarrow r + \gamma P m(\widetilde{Q}_k)$
Define $\pi_K$ as the greedy policy w.r.t. $\widetilde{Q}_K$
Output: An estimator $\widetilde{Q}_K$ of $Q^*$ and policy $\pi_K$
```

Proposition 0.35. The output $\widetilde{Q}_K$ from algorithm 2 is K-optimal and $\left\|\widetilde{Q}_K - Q^*\right\|_\infty \leqslant \gamma^K \|Q^*\|_\infty$.

Proof. See proposition 0.34. □

0.0.6 Approximation

In this section we will look at what happens if we instead use approximations of the Q-functions and T operator. This means that we are in a setting where we can somehow calculate r and TQ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, but it is hard or infeasible to represent them (or at least one of them) directly. This setting is not very well-studied in the case of a continuous state space (at least in the sources known to this writer). This is perhaps because it is considered solved by the results of theoretical Q-learning presented in the previous section. However as we have argued, this only have practical relevance when it is feasible to represent TQ. Therefore we find it relevant to consider this setting in more detail. What *is* very well-studied is a further generalized setting where T and r are assumed to be unknown, that is, one has only access to their distributions via sampling from them. We will deal with this setting in the next section. In following we present some rather simple bounding techniques which is inspired by arguments found in e.g. [**?**], together with some standard results from approximation theory on artificial neural networks and Bernstein polynomials. Throughout this section we assume (D) i.e. that we are discounting with some $\gamma \in [0, 1)$.

Let us consider any norm $\|\cdot\|$ on $(\mathcal{F},\|\cdot\|)$ where $\mathcal{F} \subseteq \mathcal{Q}$ is a subset of the space of bounded Q-functions $\mathcal{Q} = \mathcal{L}_\infty(\mathcal{S} \times \mathcal{A})$. Let $\widetilde{Q}_0$ be any Q-function which is bounded in $\|\cdot\|$. Suppose we approximate $T\widetilde{Q}_0$ by a Q-function $\widetilde{Q}_1$ to $\varepsilon_1 > 0$ precision and then approximate $T\widetilde{Q}_1$ by $\widetilde{Q}_2$ and so on. This way we get a sequence of Q-functions satisfying

$$\left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \leqslant \varepsilon_k, \forall k \in \mathbb{N}$$

First observe that

$$\begin{aligned}\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| &\leqslant \left\|T^k\widetilde{Q}_0 - T\widetilde{Q}_{k-1}\right\| + \left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\| \\ &\leqslant \gamma\left\|T^{k-1}\widetilde{Q}_0 - \widetilde{Q}_{k-1}\right\| + \left\|T\widetilde{Q}_{k-1} - \widetilde{Q}_k\right\|\end{aligned}$$

Using this iteratively we get

$$\left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \leqslant \sum_{i=1}^{k} \gamma^{k-i}\varepsilon_i := \varepsilon_{\text{approx}}(k)$$

Then we can bound

$$\begin{aligned}\left\|Q^* - \widetilde{Q}_k\right\| &\leqslant \left\|Q^* - T^k\widetilde{Q}_0\right\| + \left\|T^k\widetilde{Q}_0 - \widetilde{Q}_k\right\| \\ &\leqslant \gamma^k\left\|Q^* - \widetilde{Q}_0\right\| + \varepsilon_{\text{approx}}(k)\end{aligned}$$

These terms are called respectively the *algorithmic* error and the *approximation* error.

The algorithmic error converges exponentially, so one is often happy with this part not spending time trying to bound this tighter. The approximation error depends on our step-wise approximations. For example if $\varepsilon_i(k) = \varepsilon$ for some $\varepsilon > 0$ we easily get the bound

$$\varepsilon_{\text{approx}}(k) = \varepsilon\frac{1-\gamma^k}{1-\gamma} \leqslant \frac{\varepsilon}{1-\gamma} \tag{1}$$

If $\varepsilon_i \leqslant c\gamma^i$ we get $\varepsilon_{\text{approx}}(k) \leqslant ck\gamma^k \to 0$ as $k \to \infty$. Generally if one can show that $\varepsilon_i \to 0$ we have

Proposition 0.36. $\sum_{i-1}^{k} \gamma^{k-i}\varepsilon_i \to 0$ whenever $\varepsilon_k \to 0$ as $k \to \infty$.

Proof. Let $\varepsilon > 0$. Find N such that $\varepsilon_n \leqslant \varepsilon(1-\gamma)/2$ for all $n > N$ and find $M > N$ such that $\gamma^M \leqslant \varepsilon\gamma^N \left(\sum_{i=1}^{N} \gamma^{N-i}\varepsilon_i\right)^{-1}$. Then for all $m > M$

$$\sum_{i=1}^{m} \gamma^{m-i}\varepsilon_i \leqslant \gamma^{m-N}\sum_{i=1}^{N} \gamma^{N-i}\varepsilon_i + \sum_{i=N+1}^{m} \gamma^{m-i}\varepsilon(1-\gamma)/2 \leqslant \varepsilon/2 + \varepsilon/2 \leqslant \varepsilon$$

□

Using artifical neural networks

Setting 4. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0, 1]^w$ and $\mathcal{A}$ finite. Assume that r is continuous and P is setwise-continuous.

Definition 0.37. An **ANN** (Artificial Neural Network) with structure $(d_i)_{i=0}^{L+1} \subseteq \mathbb{N}$, activation functions $\sigma_i = (\sigma_{ij})_{j=1}^{d_i}$, where $\sigma_{ij} : \mathbb{R} \to \mathbb{R}$ are real-valued functions on $\mathbb{R}$, and weights $W_i \in M^{d_i \times d_{i-1}}$, $v_i \in \mathbb{R}^{d_i}$, $i \in [L+1]$ 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.

To clarify we have $\sigma_i(x_1, \ldots, x_{d_i}) = (\sigma_{i1}(x_1), \ldots, \sigma_{id_i}(x_{d_i}))$. $L \in \mathbb{N}_0$ is interpreted as the number of *hidden layers* and d_i is the number of neurons or nodes in layer i.

We denote the class of these networks (or functions)

$$\mathcal{DN}\left(\sigma_{ij}, (d_i)_{i=0}^{L+1}\right)$$

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

Theorem 0.38 (Universal Approximation Theorem for ANNs)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be non-constant, bounded and continuous activation function. Let $\varepsilon > 0$ and $f \in C([0,1]^w)$. Then there exists an $N \in \mathbb{N}$ and a network $F \in \mathcal{DN}(\sigma, (w, N, 1))$ with one hidden layer and activation function σ such that

$$\|F - f\|_\infty < \varepsilon$$

In other words $\bigcup_{N \in \mathbb{N}} \mathcal{DN}(\sigma, (w, N, 1))$ is dense in $C([0,1]^w)$.

Discussion of proofs. The original proof in [**?** , **?** (**?**)] is very short and elegant, but non-constructive, using the Riesz Representation and Hahn-Banach theorems to obtain a contractiction to the statement that $\bigcup_{N \in \mathbb{N}} \mathcal{DN}(\sigma, (w, N, 1))$ is dense in $C([0,1]^w)$. Furthermore it considered only *sigmoidal* activations functions, meaning that σ should satisfy

$$\sigma(x) \to \begin{cases} 0 & x \to -\infty \\ 1 & x \to \infty \end{cases}$$

This was extended in [**?** , **?** (**?**)] to the statement as presented above and their proof is constructive. □

Proposition 0.39. Consider setting 4 let and $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-constant, bounded, continuous activation function. Let $\varepsilon > 0$. Then for every $k \in \mathbb{N}$ there exists a $N \in \mathbb{N}$ and a sequence of Q-networks $(\widetilde{Q}_i)_{i=1}^k \subseteq \mathcal{DN}(\sigma, \{w|\mathcal{A}|, N, 1\})$ such that

$$\left\|T\widetilde{Q}_{i-1} - \widetilde{Q}_i\right\|_\infty < \varepsilon$$

for all $i \in [k]$. In particular

$$\left\|Q^* - \widetilde{Q}_k\right\|_\infty < \varepsilon/(1-\gamma)$$

This gives us the first method of how to approximate Q^* arbitrarily closely on continuous state spaces, in the case where it is infeasible to represent TQ directly.

Using Bernstein polynomials

We here discuss another approach using multivariate Bernstein polynomials for approximation instead of neural networks. In this case the need a slightly stronger form of continuity, namely Lipschitz continuity, to establish the bounds.

Setting 5. An MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ with $\mathcal{S} = [0,1]^w$ and $\mathcal{A}$ finite. Assume that there exists a probability measure $\mu \in \mathcal{S}$, such that $P(\cdot \mid s, a)$ has density $p(\cdot \mid s, a) : \mathcal{S} \to \mathbb{R}$ with respect to μ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. Furthermore assume that $r(\cdot, a)$, $p(s \mid \cdot, a)$ are Lipschitz with constants L_r, L_p respectively for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Definition 0.40 (Bernstein polynomial)**.** The multivariate Bernstein polynomial $B_{f,n}$ with exponents $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ approximating the function $f : [0,1]^w \to \mathbb{R}$ is defined by

$$B_{f,n}(x_1, \ldots, x_w) = \sum_{j=1}^{w} \sum_{k_j=0}^{n_j} f\left(\frac{k_1}{n_1}, \ldots, \frac{k_w}{n_w}\right) \prod_{\ell=1}^{w} \left(\binom{n_\ell}{k_\ell} x_\ell^{k_\ell} (1 - x_\ell)^{n_\ell - k_\ell} \right)$$

Notice that this a polynomial of (multivariate) degree $n_1 + \cdots + n_w$.

Theorem 0.41. Let $f : [0,1]^w \to \mathbb{R}$ be Lipschitz (see **??**) w.r.t. the standard euclidean 2-norm induced metrics on $[0,1]^w$ and $\mathbb{R}$ with constant L. Then for any $n = (n_1, \ldots, n_w) \in \mathbb{N}^w$ there exists a polynomial $B_{f,n} : [0,1]^w \to \mathbb{R}$ of degree $\leqslant \|n\|_1$ such that

1. $\left\|f - B_{f,n}\right\|_2 \leqslant \frac{L}{2}\sqrt{\sum_{j=1}^{w} \frac{1}{n_j}}$
2. $\left\|B_{f,n}\right\|_\infty \leqslant \|f\|_\infty$

Lemma 0.42. $TQ(\cdot, a)$ is Lipschitz in $\|\cdot\|_2$ with constant $L_T = (L_r + \gamma V_{\max} L_p)$ for all $a \in \mathcal{A}$ and $Q : \mathcal{S} \times \mathcal{A} \to [-V_{\max}, V_{\max}]$.

Now we can bound

Proposition 0.43.

$$\varepsilon_{\text{approx}} \leqslant \frac{L_r + \gamma V_{\max} L_p}{2(1-\gamma)} \sqrt{\sum_{j=1}^{w} \frac{1}{n_j}}$$

For example if we put $n_j = m$ for all j we get

Proposition 0.44.

$$\left\|Q^* - \tilde{Q}_k\right\| \leqslant \left\|Q^* - \tilde{Q}_0\right\| + \frac{L_r + \gamma V_{\max} L_p}{2(1-\gamma)} \sqrt{w} m^{-1/2}$$

In particular $\left\|Q^* - \tilde{Q}_k\right\|_\infty = \mathcal{O}(\gamma^{-k} + \frac{1}{\sqrt{m}})$ when using k iterations and approximating with multivariate polynomials of maximum degree $w \cdot m$.

This gives a very concrete way of constructing an arbitrarily good approximation to Q^* using polynomials.