Convergence Analysis of Deep Q-Networks

Karthik Nataraj [1]

Abstract

First I analyze close variants of Deep Q-Networks that have recently been studied, and outline important ideas, methods and performance bounds obtained in these papers, oftentimes in much greater technical detail. Second I obtain new results connecting policy and distribution assumptions used in different papers, and demonstrate improved convergence properties in (Abbeel et al., 2019) under the presence of target networks.

1. Introduction

Deep Q-learning as a technique to estimate the optimal state-action value function is relatively recent, introduced only in 2015 in (Mnih et al., 2015). Although it has enjoyed empirical success in playing games such as Atari and Go, it's theoretical foundations are less well-understood, and only recently in papers such as (Hüllermeier & Ramaswamy, 2021), (Wang et al., 2020), and (Abbeel et al., 2019) have efforts been made in this direction. By continuing to develop rigorous performance guarantees for DQN's, there is hope that it eventually can be used with more confidence in real-world settings.

2. Approach

The two main papers guiding my study were (Wang et al., 2020) and (Gu & Xu, 2020). (Wang et al., 2020) studies a simplification of the DQN, neural fitted Q-iteration, wherein the replay buffer distribution is fixed over the course of the training. It relies on a very technical analytical argument to prove an upper bound on $||Q^* - Q^{\pi_K}||_{1,\mu}$, the sum of the magnitude of distances between the optimal and returned state action value functions, weighted by an arbitrary initial distribution μ on $\mathcal{S} \times \mathcal{A}$.
(Gu & Xu, 2020) then analyzes a variant called neural Q-learning, which is essentially DQN but with no buffer and a projection step. Under a couple assumptions on the policy generating the data, it proves convergence to Q^* in time $O(1/\sqrt{T})$, T the total number of stochastic gradient descent steps. This algorithm is more similar to classical DQN, as the assumption on the policy is akin to assumptions on the distribution of the pre-generated buffer.
Given the depth of the arguments in these papers, most of my project is a synthesis of key results and their proofs. Oftentimes these results rely on earlier foundational literature on neural network function approximation (e.g. (Anthony & Bartlett, 2009), (Schmidt-Hieber)) and or function approximation with TD algorithms ((Bhandari et al., 2018), (Cai et al., 2019)) as key ingredients. Here's a brief summary of some of that previous latter work:

1. (Bhandari et al., 2018) does a finite time analysis of temporal difference learning with linear function approximation. The paper (Gu & Xu, 2020) we study in depth is similar, but for neural network function approximators.

2. (Cai et al., 2019) is also similar to (Gu & Xu, 2020) but attempts to approximate a $Q^\pi(s, a)$ instead of $Q^*(s, a)$. It also analyzes the population update setting besides the stochastic one.

Generally there is a split in prior work on TD-style algorithms, analyzing either under i.i.d. or non-i.i.d. (Markovian) assumptions on the underlying (s_t, a_t) data. Examples of the former include (Wang et al., 2020), (Liu et al., 2015), (Lakshminarayanan & Szepesvari), and (Dalal et al., 2018), while (Wang et al., 2020), (Hüllermeier & Ramaswamy, 2021), and (Gu & Xu, 2020) work under the latter assumption. It is technically more difficult to work under the non-i.i.d. assumption as updates are noisier and more biased.
Since the audience for these papers is primarily researchers in RL theory the arguments can be quite terse and unmotivated–in my exposition I prove many details that are taken for granted, and more clearly expose motivation and connect assumptions between (Wang et al., 2020) and (Gu & Xu, 2020). The report is then organized as follows: Section 3 tackles (Wang et al., 2020), Section 4 (Gu & Xu, 2020), and Section 5 contains the new results as mentioned in the abstract. Finally in Section 6 I conclude with key learnings and possible ideas for future work.

[1]ICME, Stanford University. Correspondence to: <kartnat@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

3. Neural FQI

The fitted Q-iteration algorithm is as follows:

Algorithm 1 Neural FQI

Input: MDP $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$, function class $\mathcal{F}$, (fixed) buffer distribution σ, number of iterations K, sample size n, initial state-action value function $\widetilde{Q}_0$

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

Sample n i.i.d. observations $\{s_i, a_i, r_i, s_i'\}$ from σ

Compute $Y_i = r(s_i, a_i) + \gamma \max_{a \in \mathcal{A}} \widetilde{Q}_k$

Calculate

$$\widetilde{Q}_{k+1} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(s_i, a_i))^2$$

end for

output $\widetilde{Q}_K \approx Q^*$ and corresponding greedy policy π_K

We see right away the major difference which between this algorithm and DQN: Every iteration of neural FQI a minimization problem is solved, and a new estimate of Q^* is returned. In DQN only one gradient descent step is taken per time step. This makes FQI easier to analyze in order to return an upper bound on

$$||Q^* - Q^{\pi_K}||_{1,\mu} := \sum_{(s,a)} |Q^*(s,a) - Q^{\pi_K}(s,a)| \mu(s,a),$$

so long as we can relate it to error bounds on

$$||\widetilde{Q}_k - T\widetilde{Q}_{k-1}||_\sigma := \sum_{(s,a)} (Q^*(s,a) - Q^{\pi_K}(s,a))^2 \sigma(s,a)$$

$k \in [K] := \{1, \ldots, K\}$, intermediate errors incurred in Algorithm 1. Then all that would remain would be to bound the approximation error of each $T\widetilde{Q}_k$ by our neural network function class. First the authors require an assumption on the coverage of σ over $\mathcal{S} \times \mathcal{A}$:

Assumption 1 (Concentration Coefficients)**.** *for any integer m, let $P^{\pi_m} P^{\pi_{m-1}} \cdots P^{\pi_1} \mu$ denote the distribution of (s_m, a_m) given $(s_0, a_0) \sim \mu$. Define the m-th concentration coefficient as*

$$\kappa(m; \mu, \sigma) := \sup_{\pi_1, \ldots, \pi_m} \left[\mathbb{E}_\sigma \left| \frac{d(P^{\pi_m} P^{\pi_{m-1}} \cdots P^{\pi_1} \mu)}{d\sigma} \right|^2 \right]^{1/2}.$$

Assume that there exists a constant $\phi_{\mu,\sigma} < \infty$ such that

$$(1-\gamma)^2 \sum_{m \geq 1} \gamma^{m-1} m \kappa(m; \mu, \sigma) \leq \phi_{\mu,\sigma}.$$

This assumption essentially requires that σ has sufficient coverage over $\mathcal{S} \times \mathcal{A}$. To see why, consider a countable $\mathcal{S} \times \mathcal{A}$. In this case the Radon-Nikodym derivative equals the quotient $\frac{\mu_m(s,a)}{\sigma(s,a)}$ for $\sigma(s,a) \neq 0$, μ_m being the marginal distribution of (s_m, a_m). If there is some (s,a) for which $\sigma(s,a)$ is very small while $\mu_m(s,a) \neq 0$ then this quotient blows up and the corresponding expectation does as well. On the other hand if μ_m is the same across all $(\pi_1, \ldots, \pi_m)$ and $\mu_m(s,a) = \sigma(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ where $\mu_m \neq 0$, then $\kappa(m; \mu, \sigma) = 1$ (and $\phi_{\mu,\sigma} < \infty$ clearly exists). In the latter case σ represents $\mathcal{S} \times \mathcal{A}$ near perfectly, to the extent that it is reachable after m steps of the MDP, under any nonstationary policy.

Now here is Theorem 6.1 of (Wang et al., 2020):

Theorem 1 (Error Propagation)**.** *Let Q^{π_K} be the action value function associated with the returned greedy policy π_K of algorithm 1. Under Assumption 1 we have*

$$||Q^* - Q^{\pi_K}||_{1,\mu} \leq \frac{2\phi_{\mu,\sigma}\gamma}{(1-\gamma)^2} \cdot \varepsilon_{max} + \frac{4\gamma^{K+1}}{(1-\gamma)^2} \cdot R_{max},$$

where $\varepsilon_{max} = \max_{k \in [K]} ||\widetilde{Q}_k - T\widetilde{Q}_{k-1}||_\sigma$ is the maximum single-step approximation error.

Proof. (Partial) First some notation: $\delta_k := Q_k - \widetilde{Q}_k, Q_k = T\widetilde{Q}_{k-1}, (P^\pi Q)(s,a) = \mathbb{E}[Q(s',a')|s' \sim P(\cdot|s,a), a' \sim \pi(\cdot, s')], (T^\pi Q)(s,a) = r(s,a) + \gamma(P^\pi Q)(s,a)$. Last we denote the value function upper bound $V_{\max} = R_{\max}/(1-\gamma)$. The proof then proceeds in three steps:

Step (i): First they establish a recursion relating $Q^* - \widetilde{Q}_{k+1}$ to $Q^* - \widetilde{Q}_k$. Note that

$$Q^* - \widetilde{Q}_{k+1} = Q^* - T^{\pi^*}\widetilde{Q}_k + (T^{\pi^*}\widetilde{Q}_k - T\widetilde{Q}_k) + \delta_{k+1}$$

for $k \in \{0, \ldots, K-1\}$, straight from the definition of δ_{k+1}. Letting π_k denote the greedy policy with respect to $\widetilde{Q}_k$, recall that $T\widetilde{Q}_k := T^{\pi_k}\widetilde{Q}_k \geq T^{\pi^*}\widetilde{Q}_k$. Hence

$$\begin{aligned} Q^* - \widetilde{Q}_{k+1} &= (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_k) \\ &\quad + (T^{\pi^*}\widetilde{Q}_k - T\widetilde{Q}_k) + \delta_{k+1} \\ &\leq (T^{\pi^*}Q^* - T^{\pi^*}\widetilde{Q}_k) + \delta_{k+1} \\ &= \gamma P^{\pi^*}(Q^* - \widetilde{Q}_k) + \delta_{k+1}. \end{aligned} \tag{1}$$

For a lower bound they note the following:

$$\begin{aligned} Q^* - \widetilde{Q}_{k+1} &= Q^* - T\widetilde{Q}_k + \delta_{k+1} \\ &= Q^* - T^{\pi_k}\widetilde{Q}_k + \delta_{k+1} \\ &= (TQ^* - T^{\pi_k}Q^*) + (T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k) + \delta_{k+1} \\ &\geq (T^{\pi_k}Q^* - T^{\pi_k}\widetilde{Q}_k) + \delta_{k+1} \\ &= \gamma P^{\pi_k}(Q^* - \widetilde{Q}_k) + \delta_{k+1}. \end{aligned} \tag{2}$$

With this we can inductively prove the following (which is stated without proof in (Wang et al., 2020)):

Lemma 1. *For $k, l \in \{0, 1, \ldots, K-1\}$ with $k < l$, we have*

$$Q^* - \widetilde{Q}_l \leq \sum_{i=k}^{l-1} \gamma^{l-1-i} \cdot (P^{\pi^*})^{l-1-i} \delta_{i+1} + \gamma^{l-k} \cdot (P^{\pi^*})^{l-k} (Q^* - \widetilde{Q}_k), \quad (3)$$

$$Q^* - \widetilde{Q}_l \geq \sum_{i=k}^{l-1} \gamma^{l-1-i} \cdot (P^{\pi_{l-1}} \cdots P^{\pi_{i+1}}) \delta_{i+1} + \gamma^{l-k} \cdot (P^{\pi_{l-1}} \cdots P^{\pi_k})(Q^* - \widetilde{Q}_k). \quad (4)$$

Proof. Let's consider the $\leq$ inequality first. The base case, namely $(k, l) = (0, 1)$ is trivially true (just the $k = 0$ case of equation (1). Supposing it's true for $l = l_0, 1 \leq l_0 \leq K-2$ and all $k < l_0$, we have

$$\begin{aligned} Q^* - \widetilde{Q}_{l_0+1} &\leq \gamma P^{\pi^*}(Q^* - \widetilde{Q}_{l_0}) + \delta_{l+1} \\ &\leq \sum_{i=k}^{l_0-1} \gamma^{l_0-i} (P^{\pi^*})^{l_0-i} \delta_{i+1} \\ &\quad + \gamma^{l_0+1-k} (P^{\pi^*})^{l_0+1-k} (Q^* - \widetilde{Q}_k) + \delta_{l_0+1} \\ &= \sum_{i=k}^{(l_0+1)-1} \gamma^{(l_0+1)-1-i} \cdot (P^{\pi^*})^{(l_0+1)-1-i} \delta_{i+1} \\ &\quad + \gamma^{(l_0+1)-k} (P^{\pi^*})^{(l_0+1)-k} (Q^* - \widetilde{Q}_k), \end{aligned}$$

which is just equation (3) with $l \to l_0 + 1$. So the inequality is true for $k < l_0$, and the case of $k = l_0$ is trivially true from equation (1). This completes the induction step. (4) is proven via a very similar argument. □

Of course lemma 1 only bounds errors on $Q^* - \widetilde{Q}_k$, whereas algorithm 1 returns the greedy policy π_K. We really want to understand the suboptimality of Q^{π_K} and so we need to relate these errors to $Q^* - Q^{\pi_k}$–this is what's done in the remainder of the proof. The concentration coefficients of Assumption 1 come into play in bounding terms of the form

$$\mu \left[(P^{\pi_K})^j (P^{\tau_m} P^{\tau_{m-1}} \cdots P^{\tau_1} f \right] := \widetilde{\mu}_j(f),$$

$\widetilde{\mu}_j$ the marginal distribution of $X_{j+m}, X_0 \sim \mu$. (We could see how these would arise given the terms on the RHS of 4). Then by the Cauchy-Schwarz inequality

$$\begin{aligned} \widetilde{\mu}_j(f) &\leq \left[\int_{\mathcal{S} \times \mathcal{A}} \left| \frac{d\widetilde{\mu}_J}{d\sigma} \right|^2 d\sigma \right]^{1/2} \left[\int_{\mathcal{S} \times \mathcal{A}} |f|^2 d\sigma \right]^{1/2} \\ &\leq \kappa(m + j; \mu, \sigma) \cdot ||f||_\sigma. \end{aligned}$$

□

4. Neural Q-learning

We now switch to summarize the approach in (Gu & Xu, 2020) to analyze a close variant of DQN, the neural Q-learning algorithm:

Algorithm 2 Neural Q-Learning with Gaussian Initialization

Input: learning policy π, learning rate sequence $\{\eta_t\}_{t \geq 0}, \gamma, W_l \in \mathbb{R}^{m \times m}, W_l^{(0)} \sim N(0, 1/m)$ for $l = 1, \ldots, L$
Initialization: $\theta_0 = \left(W_0^{(1)}, \ldots, W_0^{(L)}\right)$
for $t = 0, 1, \ldots, T-1$ **do**
  Sample data (s_t, a_t, r_t, s_{t+1}) from π
  $\delta_t = f(\theta_t; s_t, a_t) - (r_t + \gamma \max_{b \in \mathcal{A}} f(\theta_t; s_{t+1}, b))$
  $g_t(\theta_t) = \nabla_\theta f(\theta_t; s, a) \delta_t$
  $\theta_{t+1} = \prod_\Theta (\theta_t - \eta_t g_t(\theta_t))$
end for

Unlike in neural FQI, neural Q-learning takes into account the stochastic nature of the updates. However, this is done at the cost of a projection step, in order to keep the θ's close to θ_0. The allowable set of $\Theta = \mathbb{B}(\theta_0, \omega)$, defined as

$$\mathbb{B}(\theta_0, \omega) := \{\theta = \left(W_0^{(1)}, \ldots, W_0^{(L)}\right) : ||W_l - W_0||_F \leq \omega\}$$

for $l = 1, \ldots, L$ and fixed ω. The main theory that is utilized here is based on some work in (Cai et al., 2019), which is quite terse. Hence here I will re-do that with more of the gaps filled in:
Basically a stationary point θ^* of algorithm 2 satisfies

$$\mathbb{E}_{\mu, \pi, \mathcal{P}} \left[\delta(s, a, s'; \theta^*) \langle \nabla_\theta f(\theta^*; s, a), \theta - \theta^* \rangle \right] \geq 0,$$

where

$$\delta(s, a, s'; \theta) = f(\theta; s, a) - \left(r(s, a) + \gamma \max_{b \in \mathcal{A}} f(\theta; s', b) \right)$$

is the temporal difference error. This is because

$$\begin{aligned} &\prod_\Theta (\theta^* - \eta_t g_t(\theta^*)) \\ &= \arg\min_{\theta \in \Theta} ||\theta - (\theta^* - \eta_t g_t(\theta^*))||^2 \\ &= \arg\min_{\theta \in \Theta} \left(||\theta - \theta^*||^2 + 2\eta_t \langle \theta - \theta^*, g_t(\theta^*) \rangle \right) \\ &= \theta^*, \end{aligned}$$

as the sum in the $\arg\min$ is ≥ 0 for all θ and $= 0$ precisely at θ^*. Now consider the function family with domain $\mathcal{S} \times \mathcal{A}$:

$$\mathcal{F}_\Theta := \{f(\theta_0) + \langle \nabla_\theta f(\theta_0), \theta - \theta_0 \rangle : \theta \in \Theta\},$$

the local linearizations centered at the initialization θ_0, and note that $\langle \nabla_\theta \widehat{f}(\theta^*), \theta - \theta^* \rangle = \langle \nabla_\theta f(\theta_0), \theta - \theta^* \rangle = \widehat{f}(\theta) - \widehat{f}(\theta^*)$. Then

$$\begin{aligned} &\langle \widehat{f}(\theta^*) - \mathcal{T}\widehat{f}(\theta^*), \widehat{f}(\theta) - \widehat{f}(\theta^*) \rangle_\mu \\ &= \mathbb{E}_{\mu, \pi, \mathcal{P}} \left[\left(\widehat{f}(\theta^*) - \mathcal{T}\widehat{f}(\theta^*) \right) \left(\widehat{f}(\theta) - \widehat{f}(\theta^*) \right) \right] \\ &= \mathbb{E}_{\mu, \pi} \left[\mathbb{E}_{\mathcal{P}} [\widehat{\delta}(s, a, s'; \theta^*) \langle \nabla_\theta \widehat{f}(\theta^*; s, a), \theta - \theta^* \rangle | s, a] \right] \\ &\geq 0. \end{aligned} \quad (5)$$

Further

$$\begin{aligned}
&\prod_{\mathcal{F}_\Theta} \mathcal{T}\widehat{f}(\theta^*) \\
&= \arg\min_{g\in\mathcal{F}_\Theta} \langle g - \mathcal{T}\widehat{f}(\theta^*), g - \mathcal{T}\widehat{f}(\theta^*)\rangle_\mu \\
&= \arg\min_{g\in\mathcal{F}_\Theta} [||g - \widehat{f}(\theta^*)||^2 + \langle \widehat{f}(\theta^*) - \mathcal{T}\widehat{f}(\theta^*), g - \widehat{f}(\theta^*)\rangle_\mu \\
&+ ||\widehat{f}(\theta^*) - \mathcal{T}\widehat{f}(\theta^*)||^2 - \langle \widehat{f}(\theta^*) - \mathcal{T}\widehat{f}(\theta^*), g - \widehat{f}(\theta^*)\rangle_\mu] \\
&= \widehat{f}(\theta^*).
\end{aligned}$$

from (5). So effectively, if we substitute $f \to \widehat{f}$ in Algorithm 2 and find a stationary θ^*, then we have minimized the so-called mean-squared projected bellman error:

$$\text{MSPBE} := \mathbb{E}_{\mu,\pi,\mathcal{P}}\left[\left(\widehat{f}(\theta^*;s,a) - \prod_{\mathcal{F}_\theta}\mathcal{T}\widehat{f}(\theta^*;s,a)\right)^2\right],$$

(which $= 0$ at θ^*.). Such a θ^* does exist since

$$\begin{aligned}
&\mathbb{E}_{\mu,\pi,\mathcal{P}}\left[(\mathcal{T}\widehat{f}_1(\theta;s,a) - \mathcal{T}\widehat{f}_2(\theta;s,a))^2\right] \\
&= \gamma^2\mathbb{E}_{\mu,\pi}[\mathbb{E}_{\pi,\mathcal{P}}[\widehat{f}_1(\theta;s',a') - \widehat{f}_2(\theta;s',a')|s,a]^2] \\
&\le \gamma^2\mathbb{E}_{\mu,\pi,\mathcal{P}}[(\widehat{f}_1(\theta;s,a) - \widehat{f}_2(\theta;s,a))^2],
\end{aligned}$$

where the inequality follows from Jensen's, and so the projection onto the convex set $\mathcal{F}_\theta$ will be a contraction as well. Therefore the strategy in (Gu & Xu, 2020) is to (a) bound the l_2 normed distance between the $\widehat{f}(\theta_t)$ and $\widehat{f}(\theta^*)$, and then (b) relate these quantities to $f(\theta_t;s,a)$ and $Q^*(s,a)$. Just to give a small idea of the specifics, we have

$$\begin{aligned}
&\mathbb{E}\left[(f(\theta_T) - Q^*)^2\right] \\
&\le 3\mathbb{E}\left[(f(\theta_T) - \widehat{f}(\theta_T))^2\right] + 3\mathbb{E}\left[(\widehat{f}(\theta_T) - \widehat{f}(\theta^*))^2\right] \\
&+ 3\mathbb{E}\left[(\widehat{f}(\theta^*) - Q^*)^2\right] \qquad (6)
\end{aligned}$$

by the triangle and Cauchy-Schwartz Inequalities. The first term can be bounded by the Taylor remainder theorem, as θ_T is close to θ_0 because of the projection step. For the summand of the last term we have

$$\begin{aligned}
&|\widehat{f}(\theta^*;s,a) - Q^*(s,a)| \\
&= \left|\widehat{f}(\theta^*) - \prod_{\mathcal{F}_\Theta} Q^* + \prod_{\mathcal{F}_\Theta} Q^* - Q^*\right| \\
&= \left|\prod_{\mathcal{F}_\Theta}\mathcal{T}\widehat{f}(\theta^*) - \prod_{\mathcal{F}_\Theta}\mathcal{T}Q^* + \prod_{\mathcal{F}_\Theta} Q^* - Q^*\right| \\
&\le \left|\prod_{\mathcal{F}_\Theta}\mathcal{T}\widehat{f}(\theta^*) - \prod_{\mathcal{F}_\Theta}\mathcal{T}Q^*\right| + \left|\prod_{\mathcal{F}_\Theta} Q^* - Q^*\right| \\
&\le \gamma|\widehat{f}(\theta^*) - Q^*| + \left|\prod_{\mathcal{F}_\Theta} Q^* - Q^*\right|.
\end{aligned}$$

Therefore with a bound on $\mathbb{E}\left[(\widehat{f}(\theta_T) - \widehat{f}(\theta^*))^2\right]$ one obtains a bound on the LHS of 6 in terms of the approximation error.

5. New Results

The authors of (Gu & Xu, 2020) mention the below Assumption as key to the proof of their results:

Assumption 2. *The learning policy π and transition kernel $\mathcal{P}$ induce a Markov chain $\{s_t\}$ such that there exist constants $\lambda > 0, \rho \in (0,1)$ satisfying*

$$\sup_{s\in\mathcal{S}} d_{TV}(\mathbb{P}(s_t \in \cdot|s_0 = s), \pi') \le \lambda\rho^t,$$

for all $t \ge 0$. Here π' is the steady-state distribution of the Markov chain under policy π, and d_{TV} refers to the total variation distance between probability measures.

This is to be compared with Assumption 1, as both assumptions mandate a closeness between the data sampling distribution which is guiding the gradient descent and the data that could arise from the true MDP. One difference however is that the Markov chain of Assumption 1 is time-inhomogeneous and so may not admit a stationary distribution. In the cases where it does, however, we have the following:

Theorem 2. *Consider a finite $\mathcal{S}\times\mathcal{A}$ and suppose that there exists M such that for $m \ge M$, the* max *of the $\mathbb{E}_\sigma$ term is achieved for $\pi_1 = \pi_2 = \cdots = \pi_m$. Then Assumption 2 on the Markov chain $\{(s_t,a_t)\}$ generated from π_1 implies Assumption 1.*

Proof. Note in this case that the $\mathbb{E}_\sigma$ term is really just

$$\sum_{(s,a):\sigma(s,a)\neq 0} \frac{p_m(s,a)^2}{\sigma(s,a)},$$

where p_m is the marginal distribution of (s_m,a_m). Further $\sigma \to \pi'$ in distribution since the natural choice for σ is just the empirical distribution of observed (s,a) pairs from running π_1:

$$\sigma(s,a) = \frac{1}{m}\sum_{t\le m} 1_{\tau\sim\pi_1}\{\tau(s_t,a_t) = (s,a)\}.$$

So for some fixed $\epsilon > 0$ there exists M such that for $m \ge$

M, $\frac{\pi'(s,a)}{\sigma(s,a)} < 1+\epsilon$ for all $(s,a), \sigma(s,a) \neq 0$. Then

$$\begin{aligned}
&\sum_{(s,a):\sigma(s,a)\neq 0} \frac{p_m(s,a)^2}{\sigma(s,a)} \\
&= \sum_{(s,a)} \frac{(p_m(s,a) - \pi'(s,a) + \pi'(s,a))^2}{\sigma(s,a)} \\
&\leq \sum_{(s,a)} \frac{(\lambda\rho^m)^2}{\sigma(s,a)} + 2\lambda\rho^m \sum_{(s,a)} \frac{\pi'(s,a)}{\sigma(s,a)} \\
&+ \sum_{(s,a)} \frac{\pi'(s,a)}{\sigma(s,a)} \pi'(s,a) \\
&\leq (\lambda\rho^m)^2 \sum_{(s,a)} \frac{1}{\pi'(s,a)} + (1+\epsilon)(2\lambda\rho^m |\mathcal{S}\times\mathcal{A}| + 1) \\
&\leq C_1(\lambda\rho^m)^2 + C_2
\end{aligned}$$

for some constants $C_1, C_2 > 0$. Thus the $\kappa(m;\mu,\sigma)$ terms in

$$(1-\gamma)^2 \sum_{m\geq M} \gamma^{m-1} m\kappa(m;\mu,\sigma) \tag{7}$$

decay geometrically, implying the existence of said $\phi_{\mu,\sigma}$. □

For example, if the aforementioned π_1 is irreducible and aperiodic then the unique stationary distribution exists and Assumption 2 readily applies. It would be difficult to strengthen this theorem to accommodate situations where the $\max$ is achieved for nonstationary policies. Indeed, Assumption 2 as stated has no bearing in this setting, so it would need to be modified. Also a converse is impossible since convergence of equation 7 doesn't imply that $\kappa \to 0$ geometrically.

We now introduce theorems generalizing the analysis of (Abbeel et al., 2019) to include target networks. The authors consider the following population update:

$$\theta' = \theta + \mathbb{E}_{s,a\sim\rho}[(\mathcal{T}^*Q_\theta(s,a) - Q_\theta(s,a))\nabla_\theta Q_\theta(s,a)],$$

ρ being the distribution of experience in the replay buffer at the time of the update, and $\mathcal{T}^*$ the standard Bellman backup operator:

$$\mathcal{T}^*Q(s,a) = \mathbb{E}_{s'\sim\mathcal{P}}[r(s,a) + \gamma \max_{a'} Q^*(s',a')].$$

Then employing the Taylor expansion of Q centered at θ they derive the following expected (first-order) update, in the finite state-action space case:

$$Q_{\theta'} = Q_\theta + K_\theta D_\rho(\mathcal{T}^*Q_\theta - Q_\theta), \tag{8}$$

D_ρ the $|S||A| \times |S||A|$ diagonal matrix with entries the $\rho(s,a)$, and the $((s,a)_i,(s,a)_j)$ entry of K_θ being $\nabla_\theta Q_\theta((s,a)_i)^T \nabla_\theta Q_\theta((s,a)_j)$.

A few simplifications of this update are studied. First the tabular Q-learning operator defined as:

$$\mathcal{U}_1 Q_\theta = Q_\theta + \alpha(\mathcal{T}^*Q_\theta - Q_\theta),$$

where is proven to be a contraction when $\alpha \in (0,1)$. We have the following theorem generalizing to the presence of a target network:

Theorem 3. *In the presence of a target network $\mathcal{T}^*Q_\psi$ whose weights are updated every C steps, the $\mathcal{U}_1$ operator is a contraction with modulus $1-\alpha$ for all update steps not a positive multiple of C. Further it's fixed point is Q^*.*

Proof. To be clear, the new update operator $\mathcal{U}_1$ in the presence of a target network is

$$\mathcal{U}_1 Q_\theta = Q_\theta + \alpha(\mathcal{T}^*Q_\psi - Q_\theta).$$

One readily computes

$$\begin{aligned}
||\mathcal{U}_1Q_1 - \mathcal{U}_1Q_2||_\infty &= ||(Q_1 - \alpha Q_1) - (Q_2 - \alpha Q_2)||_\infty \\
&= (1-\alpha)||Q_1 - Q_2||_\infty,
\end{aligned}$$

and the update reverts to the old $\mathcal{U}_1$ on the updates immediately after ψ is overwritten with the current θ, every C steps. Also $\mathcal{T}^*Q_\psi = Q_\theta$ at the fixed point, and since Q_ψ is just copied from Q_θ we simply have $\mathcal{T}^*Q_\theta = Q_\theta$, implying $Q_\theta = Q^*$. □

This improves on Lemma 1 of (Abbeel et al., 2019), which shows the regular $\mathcal{U}_1$ is a contraction with modulus $1-(1-\gamma)\alpha > 1-\alpha$. Next the paper considers the following update operator:

$$\mathcal{U}_2 Q_\theta = Q_\theta + \alpha D_\rho(\mathcal{T}^*Q_\theta - Q_\theta),$$

and shows via direct computation that

$$||\mathcal{U}_2Q_1 - \mathcal{U}_2Q_2||_\infty \leq (1-(1-\gamma)\alpha\rho_{min})||Q_1 - Q_2||_\infty,$$

a contraction when $\rho_{min} > 0$ (namely the replay buffer has complete coverage) and $\alpha \in \left(0, \frac{1}{\rho(1-\gamma)}\right)$. The following theorem reveals how a target network can improve convergence in this setting:

Theorem 4. *With the target network $\mathcal{T}^*Q_\psi$, $\mathcal{U}_2$ is a contraction with modulus $1-\alpha\rho_{min} < 1-(1-\gamma)\alpha\rho_{min}$. If $\rho_{min} > 0$ and $\alpha \in \left(0, \frac{1}{\rho_{min}}\right]$ there is convergence to Q^*.*

Proof.

$$\begin{aligned}
||\mathcal{U}_2Q_1 - \mathcal{U}_2Q_2||_\infty &= ||(I - \alpha D_\rho)(Q_1 - Q_2)||_\infty \\
&\leq (1-\alpha\rho_{min})||Q_1 - Q_2||_\infty.
\end{aligned}$$

Then $0 \leq 1-\alpha\rho_{min} < 1$ mandates $\rho_{min} > 0$ and $\alpha \in \left(0, \frac{1}{\rho_{min}}\right]$, as required. Since D_ρ is invertible we also get the convergence to Q^*. □

Interestingly although we get quicker convergence with the target network, the conditions under which there is convergence are stricter, as $\frac{1}{\rho_{min}} \leq \frac{1}{\rho_{min}(1-\gamma)}$.
Finally they consider the update given by

$$\mathcal{U}_3 Q_\theta = Q_\theta + \alpha K D_\rho (\mathcal{T}^* Q_\theta - Q_\theta)$$

for K a constant symmetric, positive-definite matrix. This actually includes the case of linear function approximation for Q, namely $Q_\theta(s,a) = \theta^T \phi(s,a)$, because here the $((s,a)_i,(s,a)_j)$ entry of K_θ is just $\phi((s,a)_i)^T \phi((s,a)_j$. Their Theorem 2 states that $\mathcal{U}_3$ is a contraction when both (i) $\forall i, \alpha K_{ii}\rho_i < 1$ and (ii) $\forall i, (1+\gamma)\sum_{j\neq i}|K_{ij}|\rho_j \leq (1-\gamma)K_{ii}\rho_i$. The presence of the target network slightly relaxes the second condition:

Theorem 5. *Using the operator $\mathcal{U}_3 Q_\theta = Q_\theta + \alpha K D_\rho(\mathcal{T}^* Q_\psi - Q_\theta)$, $\mathcal{U}_3$ becomes a contraction under $\sum_{j\neq i}|K_{ij}|\rho_j < K_{ii}\rho_i$.*

Proof. Following the logic in the proof for Theorem 2 of the appendix of (Abbeel et al., 2019), we have

$$||\mathcal{U}_3 Q_1 - \mathcal{U}_3 Q_2||_\infty \leq \sum_j |\delta_{ij} - \alpha K_{ij}\rho_j|||(Q_1 - Q_2)||_\infty,$$

so we obtain a modulus of $\beta(K) = \max_i \sum_j |\delta_{ij} - \alpha K_{ij}\rho_j)| = \max_i \left(|1-\alpha K_{ii}\rho_i| + \alpha\sum_{j\neq i}|K_{ij}|\rho_j\right) = \max_i \left(1 - \alpha K_{ii}\rho_i + \alpha\sum_{j\neq i}|K_{ij}|\rho_j\right)$ (since we're also assuming condition 1 above), which is < 1 when $\sum_{j\neq i}|K_{ij}|\rho_j < K_{ii}\rho_i$, as required. □

Note that this condition is a lot looser than the previous when γ is near 1. Also if K is actually a function of θ then the above condition is sufficient for the main update operator in (8) to be contraction. Indeed, this is requiring K_θ to not generalize aggressively across different state-action pairs (namely small off-diagonal terms) *across multiple* θ, which is still quite restrictive.

6. Conclusion

Besides educating myself on theoretical guarantees for specific algorithms related to DQN, I learned about underlying methods used to make the analysis of TD-like algorithms in reinforcement learning tractable, like the usage of linear projection, studying population instead of stochastic gradients, and fixing the sampling distribution over the course of training. An interesting next step would be to see whether people have incorporated other aspects of DQN, like the ϵ-greediness and dynamic nature of the sampling distribution, into the analysis of related algorithms. Maybe these ideas could be incorporated into the analysis of neural Q-learning and/or neural FQI.

References

Abbeel, P., Achiam, J., and Knight, E. Towards characterizing divergence in deep Q-learning. *arXiv:1903.08894*, 2019.

Anthony, M. and Bartlett, P. L. Neural network learning: Theoretical foundations. *Cambridge University Press*, 2009.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1691–1692. PMLR, 06–09 Jul 2018. URL https://proceedings.mlr.press/v75/bhandari18a.html.

Cai, Q., Lee, J. D., Wang, Z., and Yang, Z. Neural temporal-difference learning converges to global optima. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/98baeb82b676b662e12a7af8ad9212f6-Paper.pdf.

Dalal, G., Mannor, S., Szörényi, B., and Thoppe, G. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. *Proceedings of the 31st Conference On Learning Theory*, 2018.

Gu, Q. and Xu, P. A finite-time analysis of Q-learning with neural network function approximation. *Proc. of Machine Learning Research*, 2020.

Hüllermeier, E. and Ramaswamy, A. Deep Q-learning: Theoretical insights from an asymptotic analysis. *IEEE Transactions on Artificial Intelligence*, 2021.

Lakshminarayanan, C. and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 09–11 Apr .

Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S., and Petrik, M. Finite-sample analysis of proximal gradient td algorithms. *Conference on Uncertainty in Artificial Intelligence*, 2015.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M. A., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C.,

Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics To appear*.

Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep Q-learning. *Proc. Learn. Dyn. Control*, pp. 486–489, 2020.