Towards Characterizing Divergence in Deep Q-Learning

arXiv 2019

Citation: 40

OpenAI, UCB

Joshua Achiam, Ethan Knight, Pieter Abbeel

Motivation

The **Deadly Triad** of DQN:

Once we put **"bootstrapping"**, **"off-policy learning"**, **"function approximation"** together, they will lead to **divergence** in DQN.

However, **the conditions under which divergence occurs are not well-understood.**

Main Ideas

Why dose DQN diverge under deadly triad?

How about analyzing DQN with NTK?

The Result of Analyzation

- The main reason why DQN diverge is **Over-generalization** and **improper(too large or too small) learning rate**.
- The **network architecture seems to affect the convergence of DQN**

Outline

- Motivation
- Main Ideas
- The Result of Analyzation
- Analyzation Setup
- NTK of DQN
- Building Intuition for Divergen with NTK
- PreQN
- Experiments

Analyzation Setup

Contraction Map

Let X be a vector space with norm $k \cdot k$, and f a function from X to X. If $\forall x, y \in X$, f satisfies

$$||f(x)-f(y)|| \leq \beta||x-y||$$

with $\beta \in [0, 1)$, then f is called a contraction map with modulus β

Banach Fixed-Point Theorem

Let f be a contraction map, $\exists x_u \ \ st \ \ f(x_u) = x_u$.

Properties

- x_u is an unique fixed-point.
- Because f is a contraction map, x_u can be obtained by the repeated application of f: for any point $x_0 \in X$, if we define a sequence of points $\{x_n\}$ such that $x_n = f(x_n - 1)$, $\lim_{n \to \infty} x_n = x$.

Bellmen Operator & Q-Function

Let $Q(s, a)$ be the Q function and $Q^*(s, a)$ be the optimal Q function.

NTK of DQN

The Bellman quation of DQN with the experience distribution ρ in replay buffer

$$Q_{k+1}(s,a) = E_{s,a\sim\rho}[Q_k(s,a) + \alpha_k(\hat{\tau}^* Q_k(s,a) - Q_k(s,a))]$$

$$\hat{\tau}^* = Q_k(s,a) = r + \gamma\, max_{a'} Q_k(s',a')$$

The TD error δ_t

$$\delta_t = \tau^* Q(s_t,a_t) - Q(s_t,a_t) = r_t + \gamma\ \max_{a'}\ Q(s_{t+1},a') - Q(s_t,a_t)$$

Update the weights

$$\theta' = \theta + \alpha E_{s,a\sim\rho}[(\tau^* Q_\theta(s,a) - Q_\theta(s,a))\ \nabla_\theta Q_\theta(s,a)]$$

NTK of DQN

The **Taylor Expansion** of Q around θ at a state-action pair $(\bar{s}, \bar{a})$.

$$Q_{\theta'}(\bar{s}, \bar{a}) = Q_{\theta}(\bar{s}, \bar{a}) + \nabla_{\theta} Q_{\theta}(\bar{s}, \bar{a})^{\top} (\theta' - \theta)$$

Combine with

$$\theta' - \theta = \alpha E_{s,a \sim \rho}[(\tau^* Q_{\theta}(s, a) - Q_{\theta}(s, a))\, \nabla_{\theta} Q_{\theta}(s, a)]$$

Thus, the Q-values before and after an update are related by:

$$Q_{\theta'}(\bar{s}, \bar{a}) = Q_{\theta}(\bar{s}, \bar{a}) + \alpha E_{s,a \sim \rho}[k_{\theta}(\bar{s}, \bar{a}, s, a)(\tau^* Q_{\theta}(s, a) - Q_{\theta}(s, a))]$$

$$k_{\theta}(\bar{s}, \bar{a}, s, a) = \nabla_{\theta} Q_{\theta}(\bar{s}, \bar{a})^{\top} \nabla_{\theta} Q_{\theta}(s, a) \qquad (9)$$

Where $k_{\theta}(\bar{s}, \bar{a}, s, a)$ is **NTK**

Building Intuition for Divergen with NTK

Theorem 1

The Q function is represented as a vector in $\mathbb{R}^{|S||A|}$, and the Q-values before and after an update are related by:

$$Q_{\theta'} = Q_\theta + \alpha K_\theta D_\rho (\tau^* Q_\theta - Q_\theta) \qquad (10)$$

where $K_\theta \in \mathbb{R}^{|S||A| \times |S||A|}$ is the matrix of entries given by the NTK $k_\theta(\bar{s}, \bar{a}, s, a)$, and D_ρ is a matrix with entries given by $\rho(s, a)$, the distribution from the replay buffer.

Consider the operator $\mathcal{U}_3$ given by

$$\mathcal{U}_3 Q = Q + \alpha K D_\rho (\tau^* Q - Q) \qquad (14)$$

Lemma 3

Under the same conditions as Theorem 1, the Q-values before and after an update are related by

$$Q_\theta = \mathcal{U}_3 Q_\theta \qquad (15)$$

Theorem 2

Let indices i, j refer to state-action pairs. **Suppose** that K and ρ satisfy the conditions:

$$\forall i,\ \alpha K_{ii}\rho_i < 1 \qquad (16)$$

$$\forall i,\ (1+\gamma)\sum_{j\neq i}|K_{ij}|\rho_j \leq (1-\gamma)K_{ii}\rho_i \qquad (17)$$

Then $\mathcal{U}_3$ is a contraction on Q in the sup norm, with fixedpoint Q^*.

Proof of Theorem 2

$$
\begin{aligned}
[\mathcal{U}_3 Q_1 - \mathcal{U}_3 Q_2]_i &= [(Q_1 + \alpha K D_\rho(\tau^* Q_1 - Q_1)) - (Q_2 + \alpha K D_\rho(\tau^* Q_2 - Q_2))]_i \\
&= [(Q_1 - Q_2) + \alpha K D_\rho((\tau^* Q_1 - Q_1) - (\tau^* Q_2 - Q_2))]_i \\
&= \sum_j \delta_{ij} [Q_1 - Q_2]_j + \alpha \sum_j K_{ij} \rho_j [(\tau^* Q_1 - Q_1) - (\tau^* Q_2 - Q_2)]_j \\
&= \sum_j (\delta_{ij} - \alpha K_{ij} \rho_j)[Q_1 - Q_2]_j + \alpha \sum_j K_{ij} \rho_j [\tau^* Q_1 - \tau^* Q_2]_j \\
&\leq \sum_j (|\delta_{ij} - \alpha K_{ij} \rho_j| + \alpha \gamma |K_{ij}| \rho_j) ||Q_1 - Q_2||_\infty
\end{aligned}
$$

Thus we can obtain a modulus as $\beta(K) = max_i \ \sum_j (|\delta_{ij} - \alpha K_{ij} \rho_j| + \alpha \gamma |K_{ij}| \rho_j)$

We'll break it up into on-diagonal and off-diagonal parts, and assume that $\alpha K_{ii}\rho_i \leq 1$:

$$\beta(K) = max_i \sum_j (|\delta_{ij} - \alpha K_{ij}\rho_j| + \alpha\gamma|K_{ij}|\rho_j)$$

$$= max_i \left((|1 - \alpha K_{ii}\rho_i| + \alpha\gamma K_{ii}\rho_i) + (1+\gamma)\alpha \sum_{j\neq i} |K_{ij}|\rho_j\right)$$

$$= max_i \left((1 - \alpha K_{ii}\rho_i + \alpha\gamma K_{ii}\rho_i) + (1+\gamma)\alpha \sum_{j\neq i} |K_{ij}|\rho_j\right)$$

$$= max_i \left(1 - (1-\gamma)\alpha K_{ii}\rho_i + (1+\gamma)\alpha \sum_{j\neq i} |K_{ij}|\rho_j\right)$$

According to Banach Fixed-Point Theorem, if $\beta(K) < 1$, $[\mathcal{U}_3 Q_1 - \mathcal{U}_3 Q_2]_i$ would converge

Thus,

$$\forall i,\ \beta(K) < 1$$

$$\forall i,\ max_i\ (1-(1-\gamma)\alpha K_{ii}\rho_i + (1+\gamma)\alpha \sum_{j\neq i} |K_{ij}|\rho_j) < 1$$

$$\forall i,\ 1-(1-\gamma)\alpha K_{ii}\rho_i + (1+\gamma)\alpha \sum_{j\neq i} |K_{ij}|\rho_j < 1$$

$$\forall i,\ (1+\gamma) \sum_{j\neq i} |K_{ij}|\rho_j < (1-\gamma)K_{ii}\rho_i$$

$$\forall i,\ \frac{(1+\gamma)}{(1-\gamma)} \sum_{j\neq i} |K_{ij}|\rho_j < K_{ii}\rho_i$$

Note that this is a quite restrictive condition, since for γ high (EX: 0.99), $(1+\gamma)/(1-\gamma)$ will be quite large, and the left hand side has a sum over all off-diagonal terms in a row.

Intuition 3

- The stability of Q-learning is **tied to the generalization properties of DQN**.
- DQNs with **more aggressive generalization (larger off-diagonal terms in K_θ)** are **less likely to demonstrate stable learning**.

Theorem 3

Consider a sequence of updates $\{\mathcal{U}_0, \mathcal{U}_1, ...\}$, with each $\mathcal{U}_i : Q \to Q$ Lipschitz continuous, with Lipschitz constant β_i , with respect to a norm $||\cdot||$. Furthermore, suppose all Ui share a common fixed-point, $\tilde{Q}$. Then for any initial point Q_0, the sequence of iterates produced by $Q_i + 1 = \mathcal{U}_i Q_i$ satisfies:

$$||\widetilde{Q} - Q_i|| \leq (\prod_{k=0}^{i-1} \beta_k)||\widetilde{Q} - Q_0||$$

Furthermore, if there is an iterate j such that $\forall k \leq j, \beta_k \in [0, 1)$, the sequence $\{\mathcal{U}_0, \mathcal{U}_1, ...\}$ converges to $\widetilde{Q}$.

Roughly speaking, this theorem says that **if you sequentially apply different contraction maps with the same fixed-point, you will attain that fixed-point which is optimal point Q^* in DQL.**

Reference

Washington University - Line Search Methods