Towards Characterizing Divergence in Deep Q-Learning

arXiv 2019

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 Analysis

- 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 Analysis
- Analysis Setup
- NTK of DQN
- Building Intuition for Divergen with NTK
- PreQN
- Experiments

Analysis 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

$$[\mathcal{U}_3 Q_1 - \mathcal{U}_3 Q_2]_i = [(Q_1 + \alpha K D_\rho(\tau^* Q_1 - Q_1)) - (Q_2 + \alpha K D_\rho(\tau^* Q_2 - Q_2))]_i$$

$$= [(Q_1 - Q_2) + \alpha K D_\rho((\tau^* Q_1 - Q_1) - (\tau^* Q_2 - Q_2))]_i$$

$$= \sum_j \delta_{ij} [Q_1 - Q_2]_j + \alpha \sum_j K_{ij} \rho_j [(\tau^* Q_1 - Q_1) - (\tau^* Q_2 - Q_2)]_j$$

$$= \sum_j (\delta_{ij} - \alpha K_{ij} \rho_j)[Q_1 - Q_2]_j + \alpha \sum_j K_{ij} \rho_j [\tau^* Q_1 - \tau^* Q_2]_j$$

$$\leq \sum_j (|\delta_{ij} - \alpha K_{ij} \rho_j| + \alpha\gamma |K_{ij}| \rho_j) ||Q_1 - Q_2||_\infty$$

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 $\alpha K_{ii}\rho_i \leq 1$
:

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

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

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

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

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.

Intuitions

- The stability and convergence 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**.
- The **architecture of network will affect to the stability and convergence of Q-learning**.
- Q-values for missing (or under-represented) state-action pairs are adjusted by **generalization with errors**. **Bootstrapping then propagates those errors** through the Q-values for all other state-action pairs.

Theorem 3

Consider a sequence of updates $\{\mathcal{U}_0, \mathcal{U}_1, ...\}$, with each $\mathcal{U}_i : Q \rightarrow Q$ 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 also optimal point Q^* in DQL.**

Preconditioned Q-Networks(PreQN)

Algorithm

Algorithm 1 PreQN (in style of DDPG)

Given: initial parameters θ, ϕ for Q, μ, empty replay buffer $\mathcal{D}$
Receive observation s_0 from environment
for $t = 0, 1, 2, ...$ **do**
  Select action $a_t = \mu_\phi(s_t) + \mathcal{N}_t$
  Step environment to get s_{t+1}, r_t and terminal signal d_t
  Store $(s_t, a_t, r_t, s_{t+1}, d_t) \rightarrow \mathcal{D}$
  if it's time to update **then**
    for however many updates **do**
      Sample minibatch $B = \{(s_i, a_i, r_i, s_i', d_i)\}$ from $\mathcal{D}$
      For each transition in B, compute TD errors:

$$\Delta_i = r_i + \gamma(1 - d_i)Q_\theta(s_i', \mu_\phi(s_i')) - Q_\theta(s_i, a_i)$$

      Compute minibatch K_θ matrix and find least-squares solution Z to $K_\theta Z = \Delta$
      Compute proposed update for Q with:

$$\theta' = \theta + \alpha_q \sum_{(s,a)\in B} Z(s,a)\nabla_\theta Q_\theta(s,a)$$

      Exponentially decay θ' towards θ until

$$\cos\left(Q_{\theta'} - Q_\theta, \mathcal{T}^* Q_\theta - Q_\theta\right) \geq \eta,$$

      then set $\theta \leftarrow \theta'$.
      Update μ with:

$$\phi \leftarrow \phi + \alpha_\mu \frac{1}{|B|} \sum_{s\in B} \nabla_\phi Q_\theta(s, \mu_\phi(s))$$

    end for
  end if
end for

We form K_θ for the minibatch, find the least-squares solution Z to

$$K_\theta Z = \tau^* Q_\theta - Q_\theta$$

For the minibatch, compute the update

$$\theta' = \theta + \alpha sum_{(s,a)\in B} Z(s,a)\nabla_\theta Q_\theta(s,a)$$

To make the

$$\cos(Q_{\theta'} - Q_\theta, \tau^* Q_\theta - Q_\theta) \leq \eta$$

Experiments

Metrics

We consider the ratio of the average off-diagonal row entry to the on-diagonal entry, R_i as '**row ratio**':

$$R_i(K) = \frac{1}{N} \frac{\sum_{j \neq i} |K_{ij}|}{K_{ii}}$$

where N is the size of the square matrix K.

The larger off-diagonal entries, the higher row ratio.

The higher row ratio, the greater generalization, but less stability and convergence

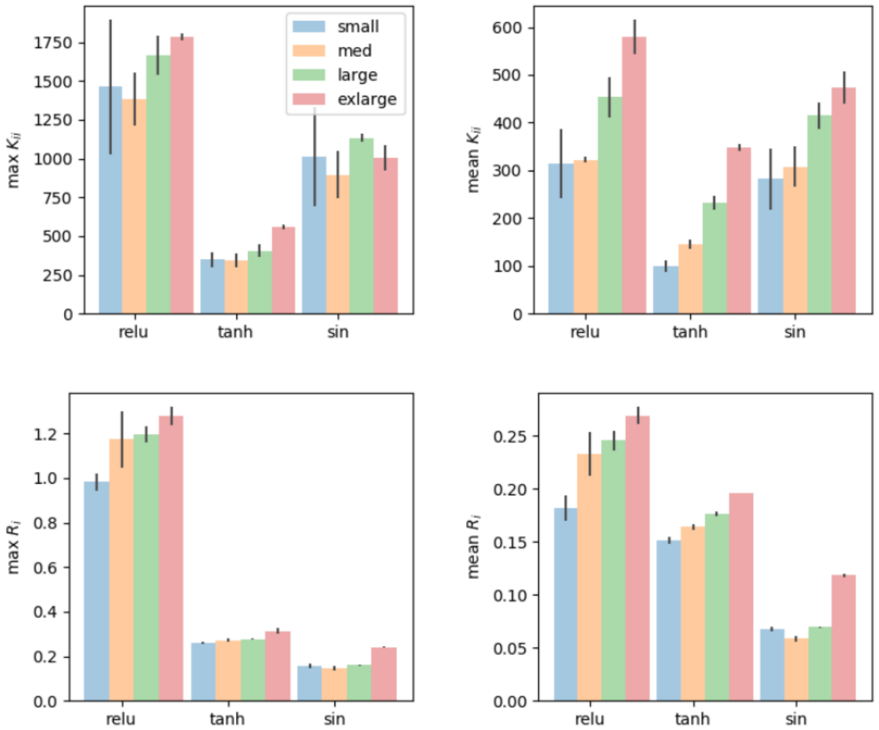

Figure 4. NTK analysis for randomly-initialized networks with various activation functions, where the NTKs were formed using 1000 steps taken by a rails-random policy in the **Ant-v2** gym environment (with the same data used across all trials). Networks are MLPs with widths of 32, 64, 128, 256 hidden units (small, med, large, exlarge respectively) and 2 hidden layers. Each bar is the average over 3 random trials (different network initializations).

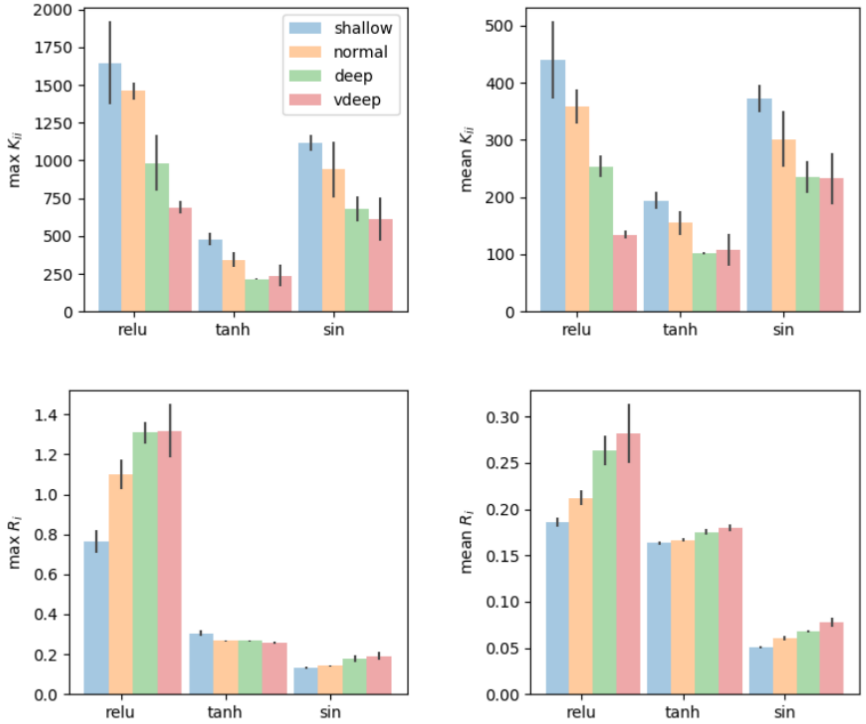

Figure 7. NTK analysis for randomly-initialized networks with various activation functions, where the NTKs were formed using 1000 steps taken by a rails-random policy in the **Ant-v2** gym environment (with the same data used across all trials). Networks are MLPs with depths of 1, 2, 3, 4 hidden layers (shallow, normal, deep, vdeep respectively) and 64 units per layer. Each bar is the average over 3 random trials (different network initializations).

- **Relu** nets commonly have the **largest on-diagonal elements and row ratio (so they should learn quickly and generalize aggressively)**
- **Sin** networks have **low off-diagonal elements and lowest row ratio.**
- Diagonal elements tend to increase with width and decrease with depth, across activation functions.
- Row ratios tend to increase with depth across activation functions, and do not clearly correlate with width.

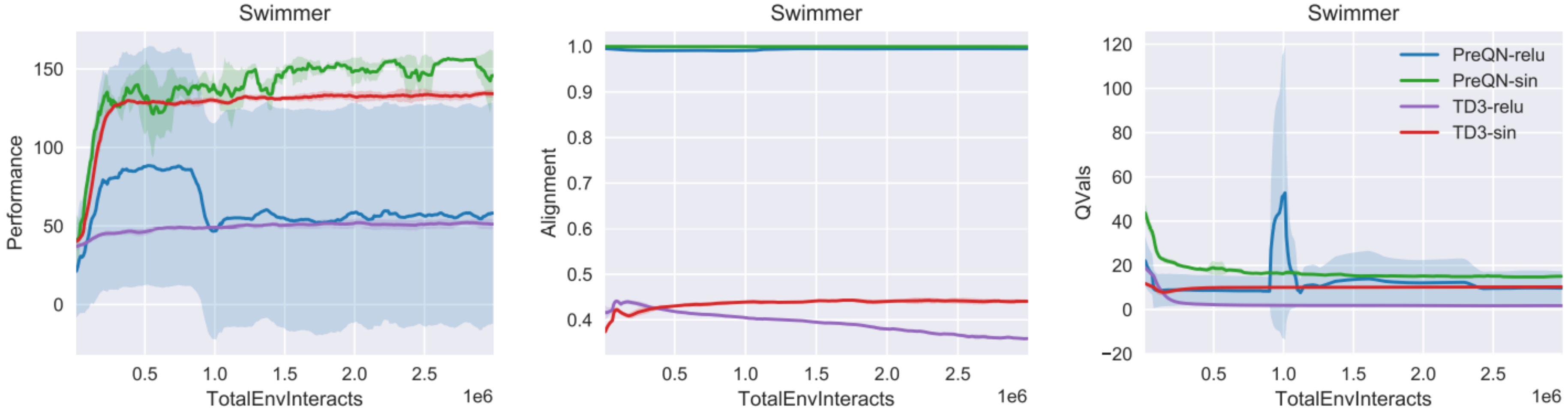

Figure 13. Comparison between PreQN and TD3 for relu and sin activation functions in the Swimmer-v2 gym environment. Results averaged over 3 random seeds.

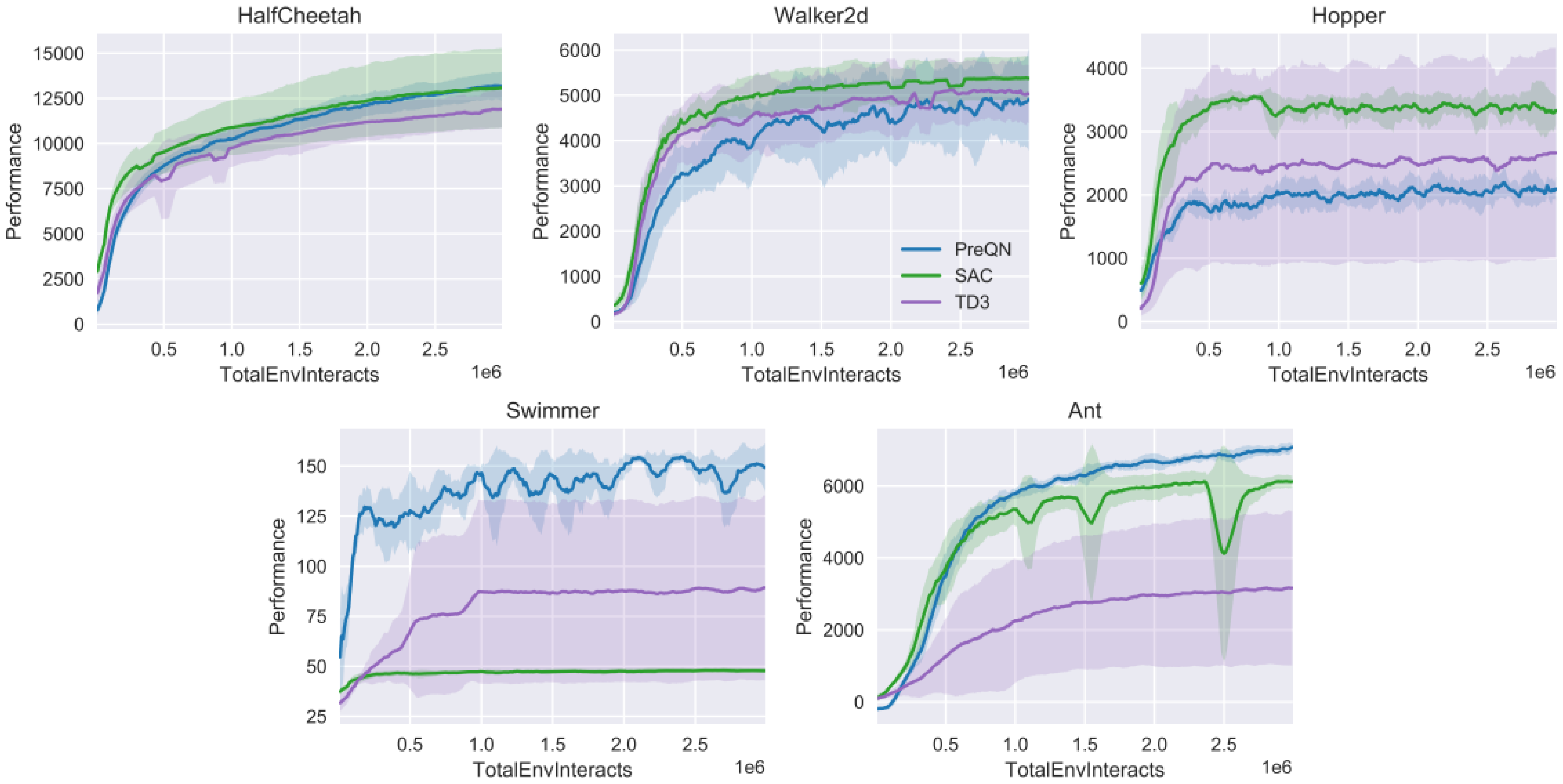

Figure 2. Benchmarking PreQN against TD3 and SAC on standard OpenAI Gym MuJoCo environments. Curves are averaged over 7 random seeds. PreQN is stable and performant, despite not using target networks. The PreQN experiments used sin activations; the TD3 and SAC experiments used relu activations.

Reference

- Washington University - Line Search Methods