MeanQ

Litian Liang · Yaosheng Xu · Stephen McAleer · Dailin Hu · Alexander Ihler · Pieter Abbeel · Roy Fox

UCI University of California, Irvine

Reducing Variance in Temporal-Difference Value Estimation via Ensemble of Deep Networks

Overview

- Ensemble mean **reduces variance** of Q value estimates
- How to sample experience to keep ensemble members **decorrelated**?

Motivation

Variance in $Q(s,a)$ hinders learning:

- $\mathrm{argmax}_a\, Q(s,a)$ is less likely **optimal**
- Can cause **instability**
  - Because similar states can appear less so
  - Target network can stabilize but adds bias
- Can add bias due to the **Jensen gap**

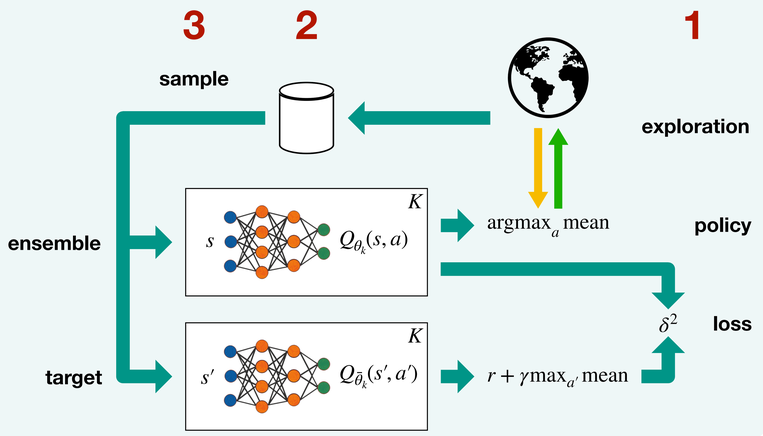
$$\mathbb{E}[\max_{a'} Q(s',a')] - \max_{a'} \mathbb{E}[Q(s',a')]$$

Related Work

- Ensemble-DQN [1], EBQL [2]
  - Train all members on the **same experience**
- Averaged-DQN [1]
  - Average snapshots of the **same learner**
- Many more (see paper)

Design Choices

1. Shared exploration policy
2. Shared replay buffer
3. Independent replay sampling

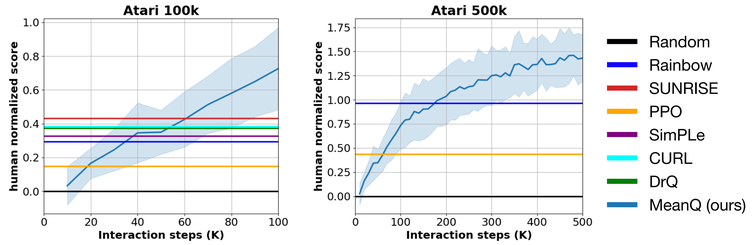


Combining with Existing Methods

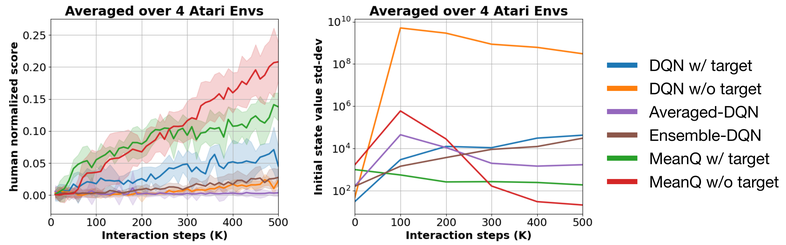
- Rainbow techniques [3]
  - Dueling nets, noisy exploration, multi-step
  - Individually prioritized experience replay
  - Distributional RL via mean distribution of Q
  - Not used: double Q-learning
- UCB exploration [4]

$$\mathrm{argmax}_a(\mathrm{mean}_k\, Q_{\theta_k}(s,a) + \lambda \mathrm{std}_k\, Q_{\theta_k}(s,a))$$

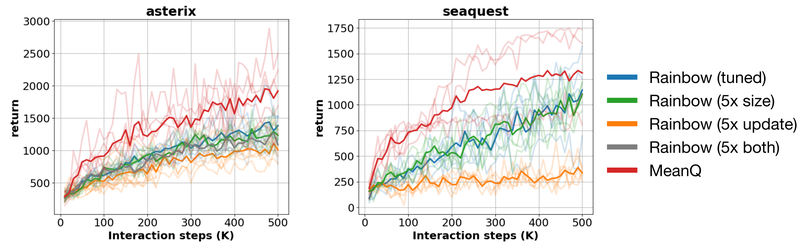
Results on Atari



No Target Network



Network Size, Update Rate



[1] *Averaged-DQN: Variance Reduction and Stabilization for Deep Reinforcement Learning*, Anschel, Baram, and Shimkin, ICML 2017

[2] *Ensemble bootstrapping for Q-learning*, Peer, Tessler, Merlis, and Meir, ICML 2021

[3] Rainbow: Combining Improvements in Deep Reinforcement Learning, Hessel et al., AAAI 2018

[4] UCB Exploration via Q-Ensembles, Chen, Sidor, Abbeel, and Schulman, arXiv 2017