Double Gumbel Q-Learning

David Yu-Tung Hui
Mila, Université de Montréal
dythui2+drl@gmail.com

Aaron Courville
Mila, Université de Montréal

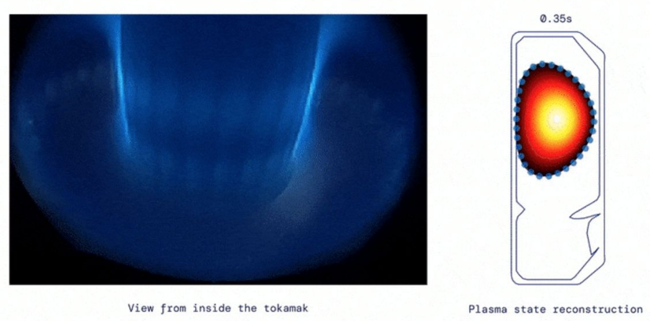
Pierre-Luc Bacon
Mila, Université de Montréal

Spotlight at NEURAL INFORMATION PROCESSING SYSTEMS 2023

RL: Reward Good Behaviors, Punish Bad Behaviors

AlphaGo
Google DeepMind, 2015 – 18

Humanoid Locomotion
Radosavovic et al., 2013



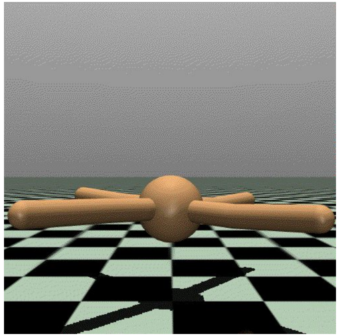
Tokamak Control
Degrave et al., 2022

This Work: Simulated Robot Control

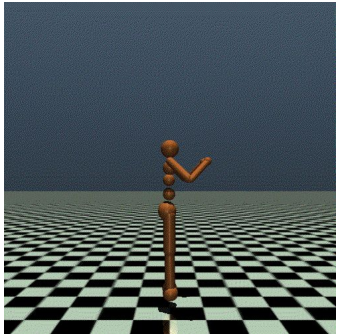
acrobot

hopper

ant



humanoid



position, velocity → motor torques

DMC: Tassa et al., 2018 [DeepMind Control Suite]() – gifs from https://github.com/facebookresearch/drqv2
MuJoCo: Brockman et al., 2016, https://github.com/openai/gym – gifs from https://gymnasium.farama.org/environments/mujoco/

DoubleGum

- Bridges gap between reinforcement learning theory and practice
- New algorithm: effective, computationally efficient, simple to implement

RL Algorithms Reinforce/Repeat Behavior that are Rewarding

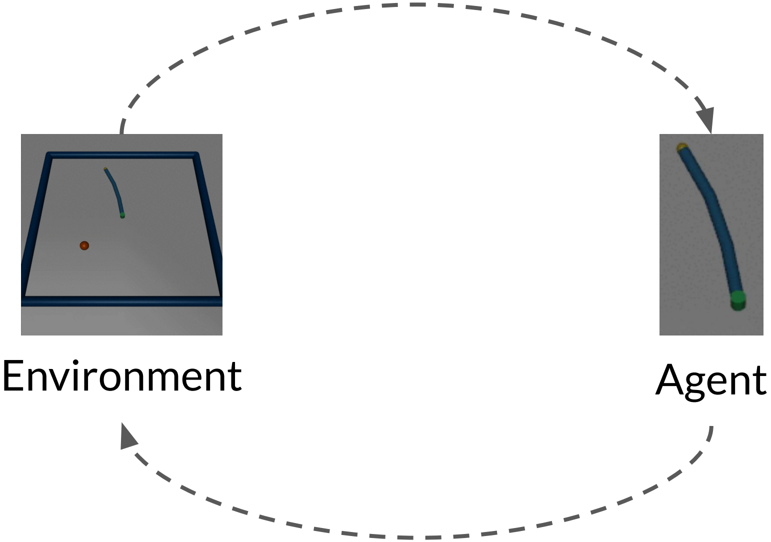


Image from https://gymnasium.farama.org/environments/mujoco/

RL Algorithms Reinforce/Repeat Behavior that are Rewarding

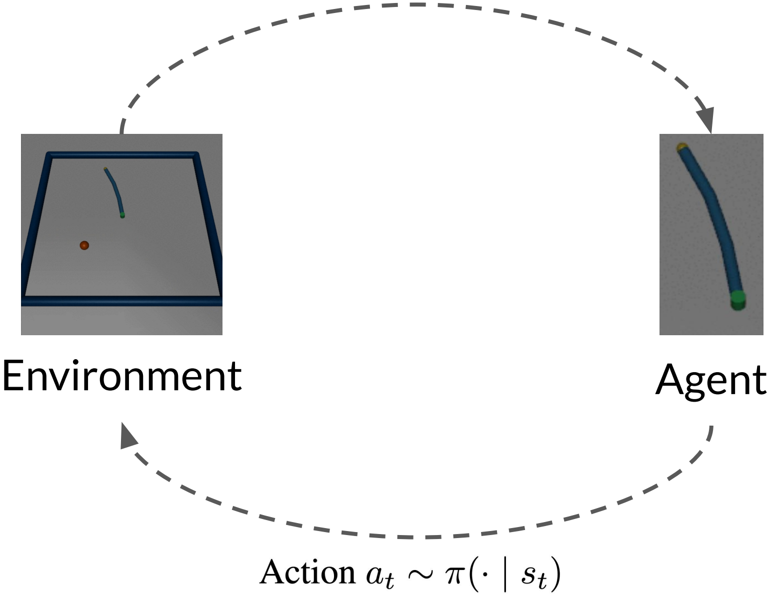
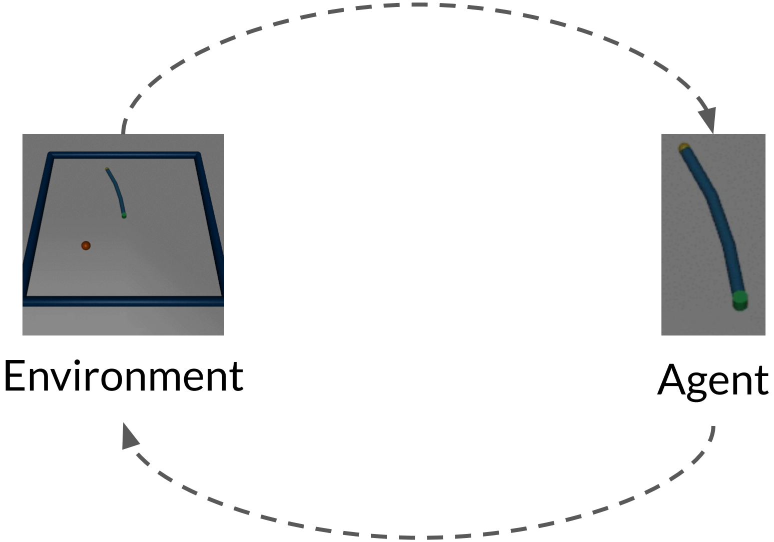


Image from https://gymnasium.farama.org/environments/mujoco/

RL Algorithms Reinforce/Repeat Behavior that are Rewarding

Transition $s_{t+1} \sim p(\cdot \mid s_t, a_t)$

Reward $r_t = r(s_t, a_t, s_{t+1})$



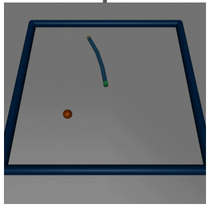
Action $a_t \sim \pi(\cdot \mid s_t)$

Image from https://gymnasium.farama.org/environments/mujoco/

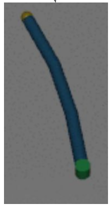
RL Algorithms Reinforce/Repeat Behavior that are Rewarding

Transition $s_{t+1} \sim p(\cdot \mid s_t, a_t)$

Reward $r_t = r(s_t, a_t, s_{t+1})$



Environment



Agent

Action $a_t \sim \pi(\cdot \mid s_t)$

Maximize Return

$$G(s_0, a_0, s_1, a_1, \ldots s_\infty) = \sum_{t=0}^{\infty} \gamma^t r_t$$

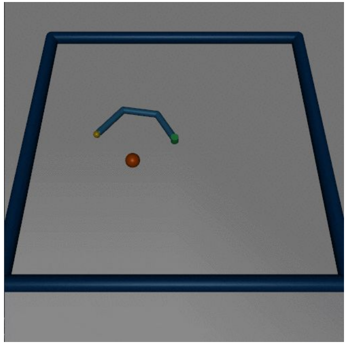


Image from https://gymnasium.farama.org/environments/mujoco/

RL Algorithms Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0, a_0, s_1, a_1, \ldots s_n)} G(s_0, a_0, s_1, a_1, \ldots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \ldots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \ldots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0,a_0,s_1,a_1,\ldots s_n)} G(s_0, a_0, s_1, a_1, \ldots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \ldots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \ldots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1}\ldots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \ldots s_n)$$

Measures quality of action a_t in s_t

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0,a_0,s_1,a_1,\dots s_n)} G(s_0, a_0, s_1, a_1, \dots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \dots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \dots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1}\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n)$$

Measures quality of action a_t in s_t

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n) \qquad \text{(Markovian Probability)}$$

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0, a_0, s_1, a_1, \ldots s_n)} G(s_0, a_0, s_1, a_1, \ldots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \ldots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \ldots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1} \ldots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \ldots s_n)$$

Measures quality of action a_t in s_t

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t, a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1}, \ldots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \ldots s_n) \quad \text{(Markovian Probability)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t, a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1}, \ldots s_n)} [r_t + \gamma G(s_{t+1}, a_{t+1}, \ldots s_n)] \quad \text{(Additive Return)}$$

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0, a_0, s_1, a_1, \ldots s_n)} G(s_0, a_0, s_1, a_1, \ldots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \ldots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \ldots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1} \ldots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \ldots s_n)$$

Measures quality of action a_t in s_t

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t, a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1}, \ldots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \ldots s_n) \quad \text{(Markovian Probability)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t, a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1}, \ldots s_n)} [r_t + \gamma G(s_{t+1}, a_{t+1}, \ldots s_n)] \quad \text{(Additive Return)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t, a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1}, a_{t+1}, \ldots s_n)} G(s_{t+1}, a_{t+1}, \ldots s_n) \right] \quad \text{(Expectation Independences)}$$

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0,a_0,s_1,a_1,\dots s_n)} G(s_0, a_0, s_1, a_1, \dots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \dots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \dots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1}\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n)$$

Measures quality of action a_t in s_t

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n) && \text{(Markovian Probability)} \\
&= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} [r_t + \gamma G(s_{t+1}, a_{t+1}, \dots s_n)] && \text{(Additive Return)} \\
&= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} G(s_{t+1}, a_{t+1}, \dots s_n) \right] && \text{(Expectation Independences)} \\
&= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} Q^\pi(s_{t+1}, a_{t+1}) \right] && \text{(Substitute } Q \text{ Definition)}
\end{aligned}
$$

RL Algorithms use a Q-Function to Maximize Expected Return

$$\max_{\pi} \mathop{\mathbb{E}}_{p_\pi(s_0,a_0,s_1,a_1,\dots s_n)} G(s_0, a_0, s_1, a_1, \dots s_n)$$

where $p_\pi(s_0, a_0, s_1, a_1, \dots s_n) = p(s_0) \prod_{t=0}^{n} p(s_{t+1} \mid s_t, a_t)\, \pi(a_t \mid s_t)$ and $G(s_0, a_0, s_1, a_1, \dots s_n) = \sum_{t=0}^{n} \gamma^t r_t$

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1}\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n)$$

Measures quality of action a_t in s_t

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} G(s_t, a_t, s_{t+1}, a_{t+1}, \dots s_n) \quad \text{(Markovian Probability)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} [r_t + \gamma G(s_{t+1}, a_{t+1}, \dots s_n)] \quad \text{(Additive Return)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} \mathop{\mathbb{E}}_{p_\pi(s_{t+1},a_{t+1},\dots s_n)} G(s_{t+1}, a_{t+1}, \dots s_n) \right] \quad \text{(Expectation Independences)}$$

$$= \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} Q^\pi(s_{t+1}, a_{t+1}) \right] \quad \text{(Substitute } Q \text{ Definition)}$$

Self-Consistency of the Q-Function

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} Q^\pi(s_{t+1}, a_{t+1}) \right]$$

$$Q^\pi(s, a) = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q^\pi(s', a') \right] \qquad \text{(Syntactic Sugar)}$$

Self-Consistency of the Q-Function

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} Q^\pi(s_{t+1}, a_{t+1}) \right]$$

$$Q^\pi(s, a) = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q^\pi(s', a') \right] \qquad \text{(Syntactic Sugar)}$$

$$\begin{aligned} Q^\star(s, a) &= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi^\star(a'|s')} Q^\star(s', a') \right] \\ &= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s', a') \right] \qquad \left(\text{because } \pi^\star(a \mid s) \text{ is determined by } \max_a Q^\star(s, a) \right) \end{aligned}$$

Self-Consistency of the Q-Function

$$Q^\pi(s_t, a_t) = \mathop{\mathbb{E}}_{p(s_{t+1}|s_t,a_t)} \left[r_t + \gamma \mathop{\mathbb{E}}_{\pi(a_{t+1}|s_{t+1})} Q^\pi(s_{t+1}, a_{t+1}) \right]$$

$$Q^\pi(s, a) = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q^\pi(s', a') \right] \qquad \text{(Syntactic Sugar)}$$

$$\begin{aligned} Q^\star(s, a) &= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi^\star(a'|s')} Q^\star(s', a') \right] \\ &= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s', a') \right] \qquad \left(\text{because } \pi^\star(a \mid s) \text{ is determined by } \max_a Q^\star(s, a) \right) \end{aligned}$$

Proof Sketch: induction with base case: $\max_a Q^\star(s_{\infty-1}, a) = \max_a \mathop{\mathbb{E}}_{p(s_\infty|\infty-1,a)} r(s_{\infty-1}, a, s_\infty)$

and inductive step: $Q^\star(s, a) = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s', a') \right]$

Deep RL Approximates Optimal Q-Function with NN

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]$$ (Bellman Optimality Equation)

Deep RL Approximates Optimal Q-Function with NN

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]$$ (Bellman Optimality Equation)

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma \max_{a'} Q_{\bar{\theta}}(s',a')\right]\right)^2$$

Deep RL Approximates Optimal Q-Function with NN

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]$$ (Bellman Optimality Equation)

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma \max_{a'} Q_{\bar{\theta}}(s',a')\right]\right)^2$$

$$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q_{\bar{\theta}}(s',a')\right]$$

$$\epsilon_{\theta,s,a} \sim \mathcal{N}(0,\sigma)$$

Note after presenting: Make it clearer that this MLE interpretation of the MSE loss applied to Q-learning is my interpretation (from many), and that this is not presented in the papers that applied MSE to Q-Learning [Riedmiller, 2005] and [Ernst, 2005]

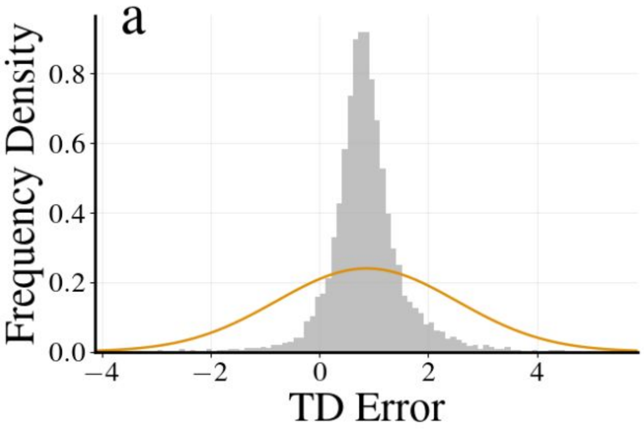
Deep RL Approximates Optimal Q-Function with NN

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]$$ (Bellman Optimality Equation)

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma \max_{a'} Q_{\bar\theta}(s',a')\right]\right)^2$$

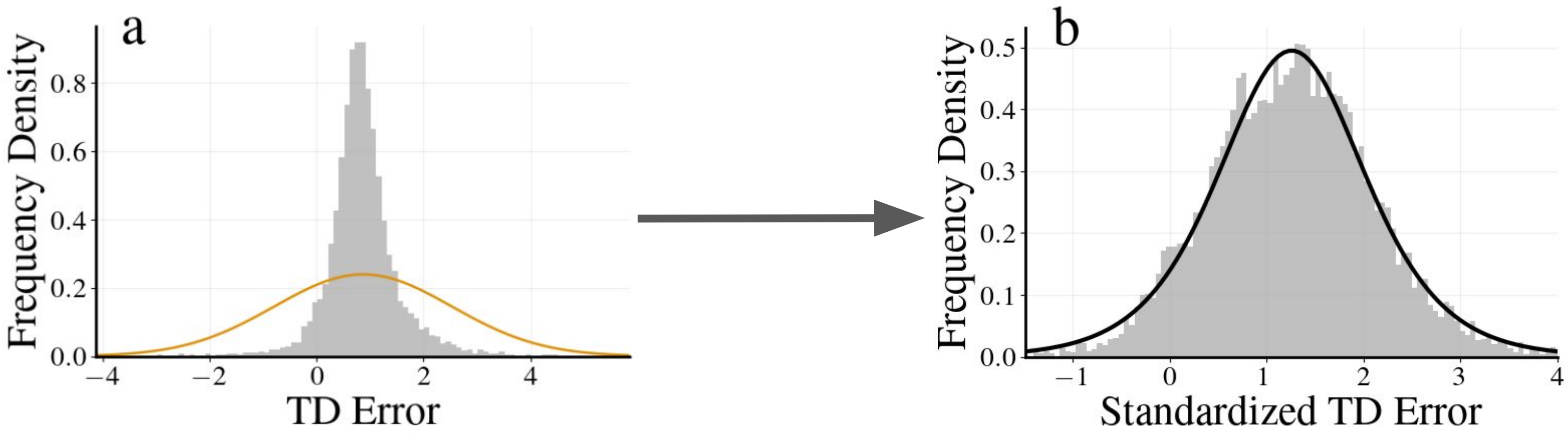
$$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q_{\bar\theta}(s',a')\right]$$

$$\epsilon_{\theta,s,a} \sim \mathcal{N}(0,\sigma)$$



Modelling Assumption of MSE does not match empirical behavior!

Deep Q-Learning has an inaccurate noise model
We present a theoretically-grounded noise model leading to better performance



Standard Noise Model

Our Noise Model

1. Introduction
2. **Derivation**
3. Results

1. Introduction

2. Derivation
    a. **Of noise model**
    b. Of algorithm

3. Results

Noise Model Derivation Outline

$$\mathrm{L}\left(Q_\theta(s,a), Q^\star(s,a)\right) \qquad \text{(Objective)}$$

Noise Model Derivation Outline

$$\begin{aligned} & \mathrm{L}\left(Q_\theta(s,a), Q^\star(s,a)\right) && \text{(Objective)} \\ & = \mathrm{L}\left(Q_\theta(s,a), \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]\right) && \text{(Bellman Equation)} \end{aligned}$$

Noise Model Derivation Outline

$$
\begin{aligned}
& \mathrm{L}\left(Q_\theta(s,a), Q^\star(s,a)\right) && \text{(Objective)} \\
&= \mathrm{L}\left(Q_\theta(s,a), \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]\right) && \text{(Bellman Equation)} \\
&= \mathrm{L}_1\left(Q_\theta(s,a), \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q_\theta(s',a')\right]\right) && \text{(Bootstrapping)}
\end{aligned}
$$

Two Components of the DDPG Baseline Algorithm

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right] \longrightarrow$$

One Neural Network:

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma \max_{a'} Q_{\bar{\theta}}(s',a')\right]\right)^2$$

Two Components of the DDPG Baseline Algorithm

$$Q^\star(s,a) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]$$

One Neural Network:

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma \max_{a'} Q_{\bar\theta}(s',a')\right]\right)^2$$

Two Neural Networks:

Critic learns Q-Function:

$$\min_\theta \sum_i \left(Q_\theta(s_i,a_i) - \mathop{\mathbb{E}}_{p(s'|s_i,a_i)}\left[r + \gamma Q_{\bar\theta}(s',\pi_\phi(s'))\right]\right)^2$$

Actor learns to maximize Q-Function:

$$\max_\phi \sum_i Q_\theta(s_i,\pi_\phi(s_i))$$

DDPG: Deep Deterministic Policy Gradients [Lillicrap et al., 2015]

Noise Model Derivation Outline

$$
\begin{aligned}
&\mathrm{L}\left(Q_\theta(s,a), Q^\star(s,a)\right) && \text{(Objective)}\\
&= \mathrm{L}\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]\right) && \text{(Bellman Equation)}\\
&= \mathrm{L}_1\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \max_{a'} Q_\theta(s',a')\right]\right) && \text{(Bootstrapping)}\\
&= \mathrm{L}_2\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \underset{\pi(a'|s')}{\mathbb{E}} Q_\theta(s',a')\right]\right) && \text{(Actor)}
\end{aligned}
$$

Noise Model Derivation Outline

$$
\begin{aligned}
&\mathrm{L}\left(Q_\theta(s,a), Q^\star(s,a)\right) && \text{(Objective)}\\
&= \mathrm{L}\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \max_{a'} Q^\star(s',a')\right]\right) && \text{(Bellman Equation)}\\
&= \mathrm{L}_1\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \max_{a'} Q_\theta(s',a')\right]\right) && \text{(Bootstrapping)}\\
&= \mathrm{L}_2\left(Q_\theta(s,a), \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \underset{\pi(a'|s')}{\mathbb{E}} Q_\theta(s',a')\right]\right) && \text{(Actor)}
\end{aligned}
$$

Quantify Noise Induced by Each Approximation

(Objective)

(Bellman Equation)

(Bootstrapping)

(Actor)

(Objective from noise model)

(Bellman Equation)

(Bootstrapping)

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a)$ $\quad$ $g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot))$ $\quad$ (Objective from noise model)

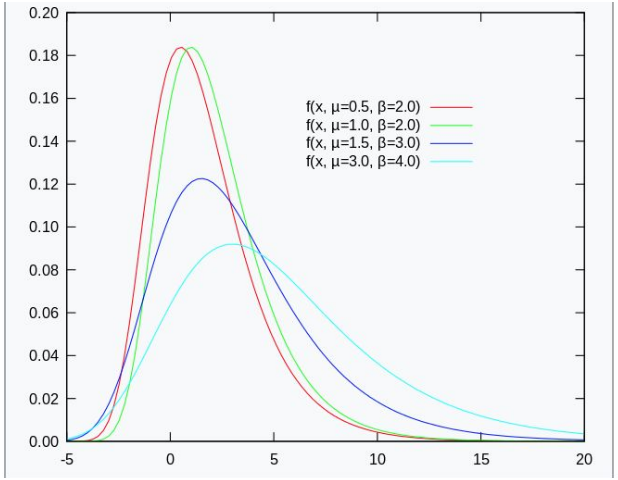
(Bellman Equation)

(Bootstrapping)

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$



Notation	$\mathrm{Gumbel}(\mu, \beta)$
Parameters	μ, location (real) $\beta > 0$, scale (real)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\beta} e^{-(z+e^{-z})}$ where $z = \frac{x-\mu}{\beta}$
CDF	$e^{-e^{-(x-\mu)/\beta}}$

Extreme Value Theorem: $\max_i z_i \sim \mathcal{G}(\alpha, \beta) \ , \ z_i \sim \text{Noise}$

(if z unbounded)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

[Thrun and Schwarz, 1993]: $Q_\theta(s,a) = Q^\star(s,a) + z_{\theta,s,a}$, $\quad z_{\theta,s,a} \sim \text{Noise}$

Ours: $Q_\theta(s,a) = Q^\star(s,a) - g_{a,\theta}(s) + g_\theta(s)$, $\quad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot))$

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s', a')\right] \qquad \text{(Bellman Equation)}$$

(Bootstrapping)

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$
\begin{aligned}
Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) &= Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) && \text{(Objective from noise model)} \\
&= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s',a') \right] && \text{(Bellman Equation)} \\
&= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} [Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')] \right] && \text{(Bootstrapping)}
\end{aligned}
$$

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')\right]}\right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s')\right]} = \max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s')\right] - g_\theta(s') \qquad \text{(Independence of } a'\text{)}$$

Typo: should be g_{\theta, a'} here

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')\right]}\right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s')\right]} = \max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s')\right] - g_\theta(s') \qquad \text{(Independence of } a'\text{)}$$

Typo: should be g_{\theta, a'} here

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) + g'_\theta(s') - g_\theta(s') \qquad \text{(Gumbel Max-Stability)}$$

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')\right]}\right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s')\right]} = \max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s')\right] - g_\theta(s') \qquad \text{(Independence of } a'\text{)}$$

Typo: should be g_{\theta, a'} here

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) + g'_\theta(s') - g_\theta(s') \qquad \text{(Gumbel Max-Stability)}$$

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) \qquad \text{(Distributions Cancel)}$$

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s', a') \right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \boxed{\max_{a'} [Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')]} \right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'} [Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s')]} = \max_{a'} [Q_\theta(s',a') + g_{\theta,a}(s')] - g_\theta(s') \qquad \text{(Independence of } a')$$

Typo: should be g_{\theta, a'} here

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) + g'_\theta(s') - g_\theta(s') \qquad \text{(Gumbel Max-Stability)}$$

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) \qquad \text{(Distributions Cancel)}$$

$$= \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, \mathbb{C}\left[\pi_\phi \,||\, p_\theta\right] \qquad \text{(Soft } Q\text{-Learning Identity)}$$

where $p_\theta(a \mid s) = \frac{1}{Z} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right)$

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \max_{a'} Q^\star(s', a') \right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \boxed{\max_{a'} \left[Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s') \right]} \right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'} \left[Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s') \right]} = \max_{a'} \left[Q_\theta(s',a') + g_{\theta,a}(s') \right] - g_\theta(s') \qquad \text{(Independence of } a'\text{)}$$

Typo: should be g_{\theta, a'} here

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')} \right) + g'_\theta(s') - g_\theta(s') \qquad \text{(Gumbel Max-Stability)}$$

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')} \right) \qquad \text{(Distributions Cancel)}$$

$$= \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s') \, \mathbb{C}\left[\pi_\phi \,||\, p_\theta \right] \qquad \text{(Soft } Q\text{-Learning Identity)}$$

where $\quad p_\theta(a \mid s) = \frac{1}{Z} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')} \right)$

(Actor)

Assumption: Two Heteroscedastic Gumbel Noise Sources

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = Q^\star(s,a) \qquad g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0,\beta(\cdot)) \qquad \text{(Objective from noise model)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q^\star(s',a')\right] \qquad \text{(Bellman Equation)}$$

$$= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a'}(s') - g_\theta(s')\right]}\right] \qquad \text{(Bootstrapping)}$$

$$\boxed{\max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s') - g_\theta(s')\right]} = \max_{a'}\left[Q_\theta(s',a') + g_{\theta,a}(s')\right] - g_\theta(s') \qquad \text{(Independence of } a'\text{)}$$

Typo: should be g_{\theta, a'} here

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) + g'_\theta(s') - g_\theta(s') \qquad \text{(Gumbel Max-Stability)}$$

$$= \beta(s') \log \sum_{a'} \exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right) \qquad \text{(Distributions Cancel)}$$

$$= \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, \mathbb{C}\left[\pi_\phi \,||\, p_\theta\right] \qquad \text{(Soft } Q\text{-Learning Identity)}$$

where $p_\theta(a \mid s) = \frac{1}{Z}\exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right)$

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, \mathbb{C}\left[\pi_\phi \,||\, p_\theta\right]\right] \qquad \text{(Actor)}$$

DoubleGum Noise Model

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \;||\; p_\theta\right]\right]$$

where $p_\theta(a \mid s) = \frac{1}{Z}\exp\left(\frac{Q_\theta(s',a')}{\beta(s')}\right)$

$$g_{a,\theta}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0, \beta(\cdot))$$

1. Introduction

2. Derivation
   a. Of noise model
   b. **Of algorithm**

3. Results

Noise Model -> Loss Function

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, \mathbb{C}\left[\pi_\phi \;||\; p_\theta\right]\right]$$

Noise Model -> Loss Function

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \mid\mid p_\theta\right]\right]$$

$$\begin{aligned}
\text{LHS} &= Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) \;\;, \qquad & g_{\theta,a}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0,\beta(\cdot)) \\
&= Q_\theta(s,a) + l_{\theta,a}(s) \;\;, \qquad & l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0,\beta(\cdot))
\end{aligned}$$

Logistic Distribution

Noise Model -> Loss Function

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \mid\mid p_\theta\right]\right]$$

$$\begin{aligned}\text{LHS} &= Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s)\ \ , && g_{\theta,a}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0,\beta(\cdot)) \\ &= Q_\theta(s,a) + l_{\theta,a}(s)\ \ , && l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0,\beta(\cdot))\end{aligned}$$

Logistic Distribution

$$\begin{aligned}\text{RHS} &= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \mid\mid p_\theta\right]\right] \\ &\approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right]\end{aligned}$$

Noise Model -> Loss Function

$$Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s) = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \mid\mid p_\theta\right]\right]$$

$$\begin{aligned}\mathrm{LHS} &= Q_\theta(s,a) + g_{\theta,a}(s) - g_\theta(s)\ , \qquad && g_{\theta,a}(\cdot), g_\theta(\cdot) \overset{\text{iid}}{\sim} \mathcal{G}(0,\beta(\cdot)) \\ &= Q_\theta(s,a) + l_{\theta,a}(s)\ , && l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0,\beta(\cdot))\end{aligned}$$

Logistic Distribution

$$\begin{aligned}\mathrm{RHS} &= \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,\mathbb{C}\left[\pi_\phi \mid\mid p_\theta\right]\right] \\ &\approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,c\right]\end{aligned}$$

Scalar hyperparameter c

$$\mathrm{LHS} = \mathrm{RHS}$$

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,c\right]$$

The DoubleGum Algorithm

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,c\right] \qquad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

1. c hyperparameter determined at beginning of training and fixed
2. Learn \beta and q using generalized method of moments

The DoubleGum Algorithm

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c \right] \qquad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

1. c hyperparameter determined at beginning of training and fixed
2. Learn \beta and q using generalized method of moments

$$\begin{aligned} & l_{\theta,a}(s) \;\;,\;\; l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot)) \\ \approx \;\; & n_{\theta,a}(s) \;\;,\;\; n_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \beta(\cdot) \frac{\pi}{\sqrt{3}}\right) \end{aligned}$$

The DoubleGum Algorithm

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \boxed{\mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right]} \qquad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

1. c hyperparameter determined at beginning of training and fixed
2. Learn \beta and q using generalized method of moments

$$\begin{aligned} & l_{\theta,a}(s) \;\;,\;\; l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot)) \\ \approx\;\; & n_{\theta,a}(s) \;\;,\;\; n_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \beta(\cdot)\frac{\pi}{\sqrt{3}}\right) \end{aligned}$$

Moment-Matching loss function:

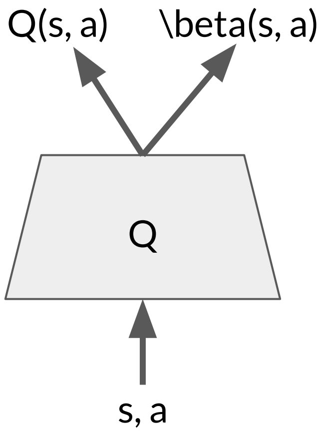
$$\min_\theta \left[\log \beta_\theta(s,a)\frac{\pi}{\sqrt{3}} + \frac{3}{\pi^2}\frac{1}{\beta_\theta(s,a)^2}\left(Q_\theta(s,a) - \boxed{y(s,a)}\right)^2\right]$$

The DoubleGum Algorithm

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \boxed{\mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right]} \qquad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

1. c hyperparameter determined at beginning of training and fixed
2. Learn \beta and q using generalized method of moments

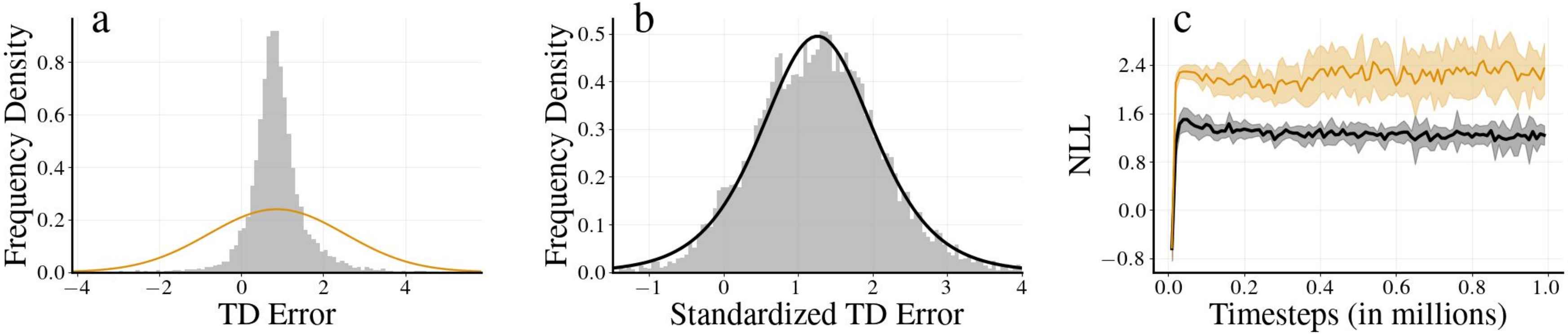
$$\begin{aligned} & l_{\theta,a}(s) \;\; , \;\; l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot)) \\ \approx \;\; & n_{\theta,a}(s) \;\; , \;\; n_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \beta(\cdot)\frac{\pi}{\sqrt{3}}\right) \end{aligned}$$



Moment-Matching loss function:

$$\min_\theta \left[\log \beta_\theta(s,a)\frac{\pi}{\sqrt{3}} + \frac{3}{\pi^2}\frac{1}{\beta_\theta(s,a)^2}\left(Q_\theta(s,a) - \boxed{y(s,a)}\right)^2\right]$$

Empirical Validity of DoubleGum Noise Model



Hetero-Logistic (Ours) $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right]$$

Homo-Normal (DDPG) $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \max_{a'} Q_\theta(s',a')\right]$$

The DoubleGum Algorithm

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\, c \right]$$

where $l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$

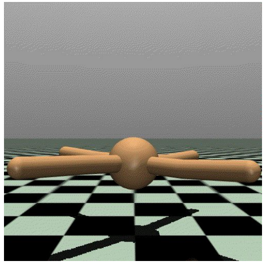
1. c fixed hyperparameter determined at beginning of training
2. Learn \beta and q using generalized method of moments

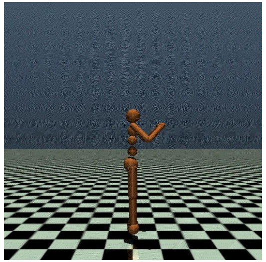
1. Introduction
2. Derivation
3. **Results**

Four Simulated Robotics Suites

DeepMind Control (DMC)
Locomotion

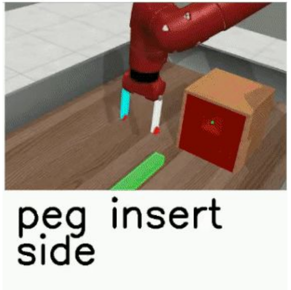
Movement, Joints + Control (MuJoCo)
Locomotion

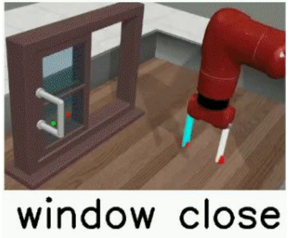




Meta-World

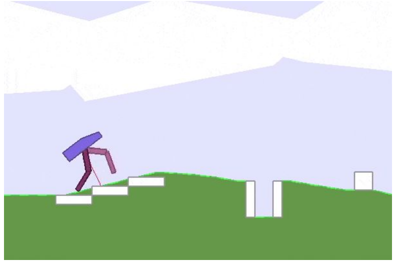
Manipulation





Box2D

Locomotion



DMC: Tassa et al., 2018 – gifs from https://github.com/facebookresearch/drqv2
MuJoCo: Brockman et al., 2016 – gifs from https://gymnasium.farama.org/environments/mujoco/
Meta-World: Yu et al., 2019
Box2D: https://gymnasium.farama.org/environments/box2d/

Baselines

DoubleGum: $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\,c\right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

Baselines

DoubleGum: $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

DDPG: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a')\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

Baselines

DoubleGum: $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, c \right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

DDPG: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') \right] \quad , \qquad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

TD3: $$Q_{\theta_i}(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)} \left[r + \gamma \min_i \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_{\theta_i}(s',a') \right] \quad , \qquad i = \{1,2\} \;\; , \;\; \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

Baselines

DoubleGum: $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

DDPG: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a')\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

TD3: $$Q_{\theta_i}(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \min_i \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_{\theta_i}(s',a')\right] \quad , \quad i = \{1,2\} \ , \ \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

SAC: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \mathbb{H}[\pi_\phi]\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$$

Baselines

DoubleGum: $$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \beta(s')\, c\right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0,\beta(\cdot))$$

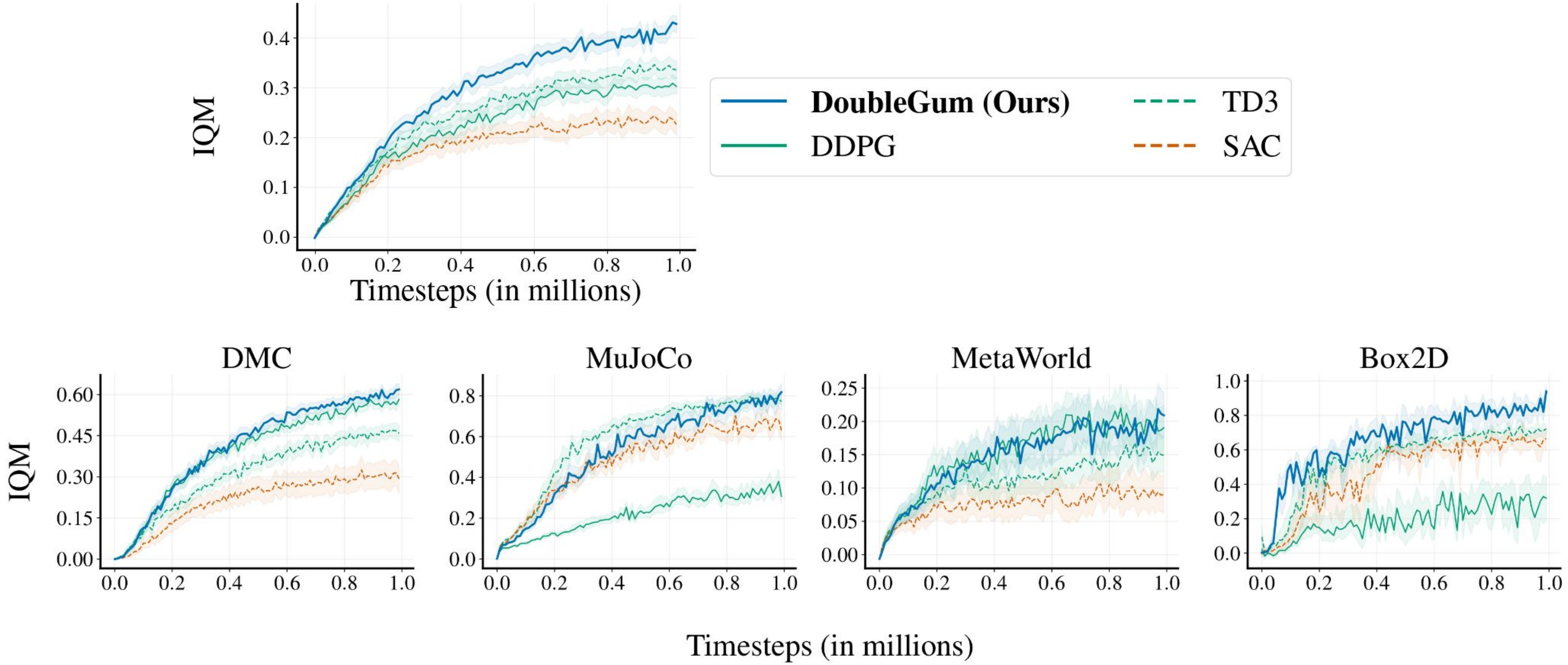
DDPG: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a')\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0,\beta(\cdot))$$

TD3: $$Q_{\theta_i}(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \min_i \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_{\theta_i}(s',a')\right] \quad , \quad i = \{1,2\} \;,\; \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0,\beta(\cdot))$$

SAC: $$Q_\theta(s,a) + \epsilon_{\theta,s,a} = \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi_\phi(a'|s')} Q_\theta(s',a') + \mathbb{H}[\pi_\phi]\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0,\beta(\cdot))$$

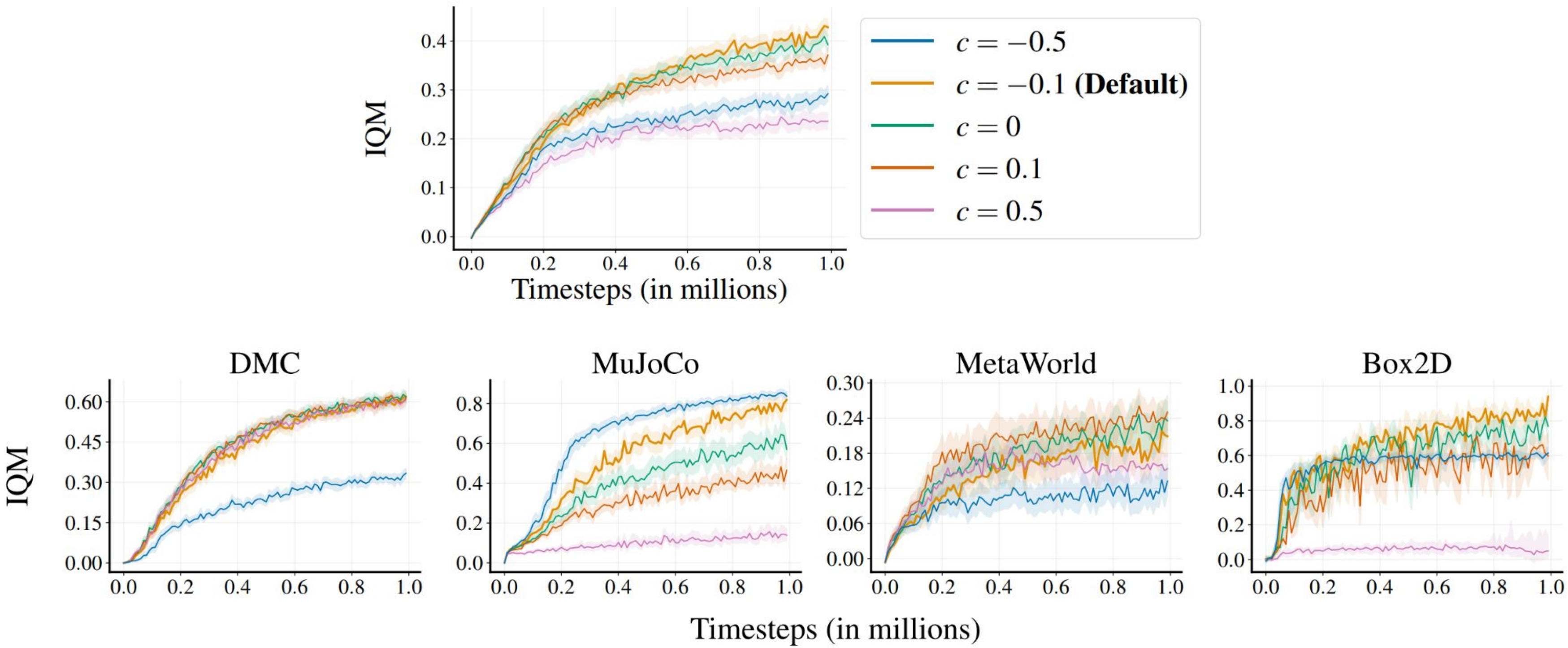
Evaluation mode of DoubleGum: hyperparameter c=-0.1 fixed across all tasks

Benchmark on 33 Continuous Control Tasks, 4 Suites



IQM: InterQuartile Mean +- 95% stratified bootstrap CIs. An aggregate metric, from [Agarwal et al., 2021]

Varying hyperparameter c



Varying c changes pessimism/optimism of target

$$Q_\theta(s,a) + l_{\theta,a}(s) \approx \mathop{\mathbb{E}}_{p(s'|s,a)}\left[r + \gamma \mathop{\mathbb{E}}_{\pi(a'|s')} Q_\theta(s',a') + \beta(s')\,c\right] \quad \text{where} \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$$

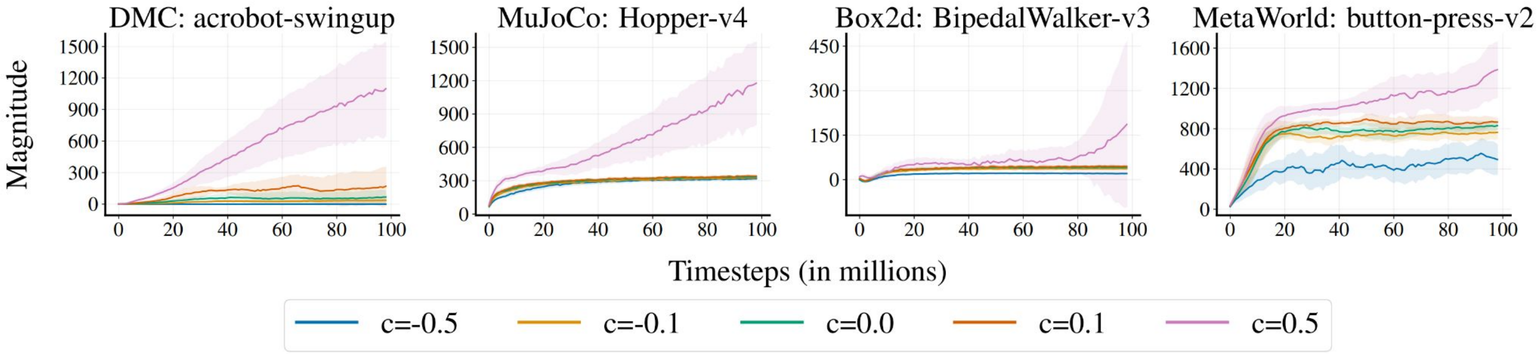


Figure 2: The effect of changing pessimism factor c on the target Q-value in continuous control

Baselines: adjusting pessimism per suite

DoubleGum: $Q_\theta(s,a) + l_{\theta,a}(s) \approx \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \underset{\pi_\phi(a'|s')}{\mathbb{E}} Q_\theta(s',a') + \beta(s')\,c\right] \quad , \quad l_{\theta,a}(\cdot) \overset{\text{iid}}{\sim} \mathcal{L}(0, \beta(\cdot))$

DDPG: $Q_\theta(s,a) + \epsilon_{\theta,s,a} = \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \underset{\pi_\phi(a'|s')}{\mathbb{E}} Q_\theta(s',a')\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$

TD3: $Q_{\theta_i}(s,a) + \epsilon_{\theta,s,a} = \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \min_i \underset{\pi_\phi(a'|s')}{\mathbb{E}} Q_{\theta_i}(s',a')\right] \quad , \quad i = \{1,2\} \;,\; \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$

SAC: $Q_\theta(s,a) + \epsilon_{\theta,s,a} = \underset{p(s'|s,a)}{\mathbb{E}}\left[r + \gamma \underset{\pi_\phi(a'|s')}{\mathbb{E}} Q_\theta(s',a') + \mathbb{H}[\pi_\phi]\right] \quad , \quad \epsilon_{\theta,s,a} \overset{\text{iid}}{\sim} \mathcal{N}(0, \beta(\cdot))$

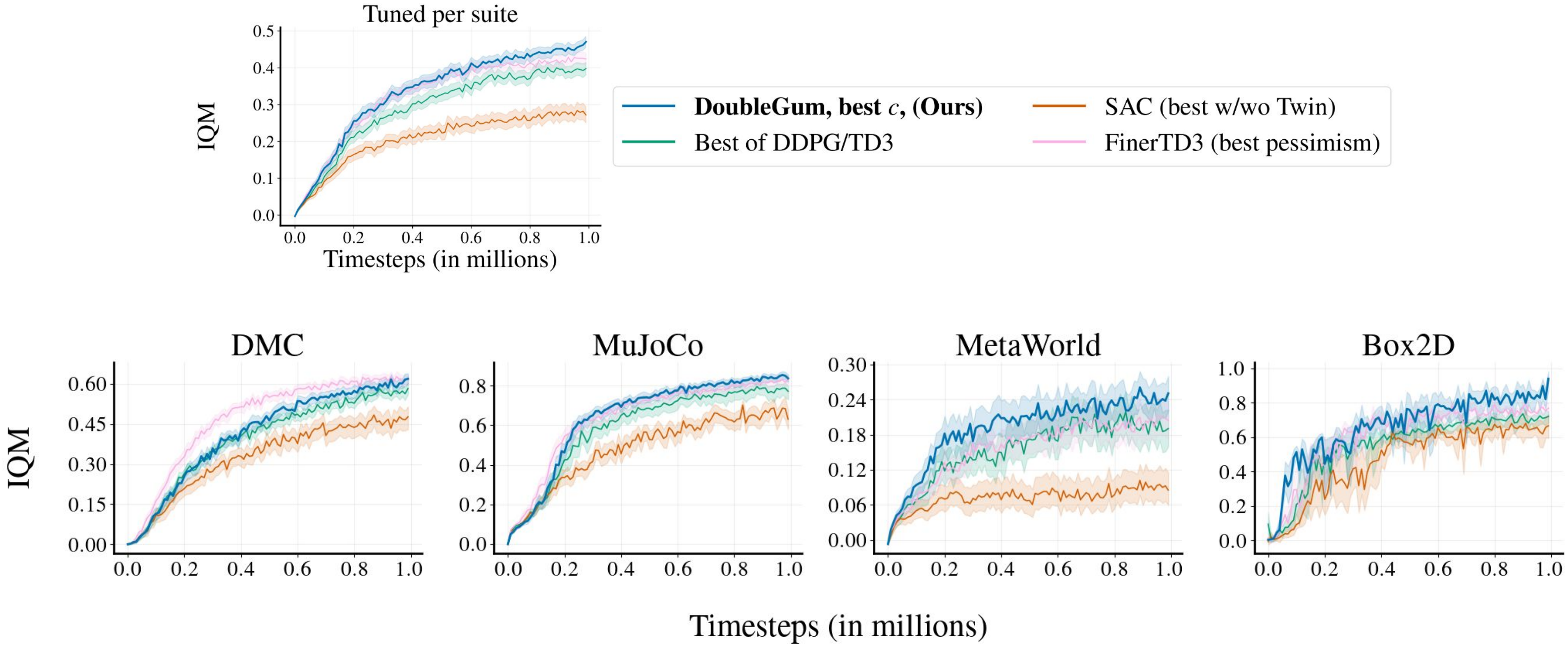
Typo: should be \beta \mathbb{H} here

Best of DDPG/TD3

Apply Twin Networks to SAC

FinerTD3: i = {1, 2, 3, 4, 5}, select j^{th} smallest value

Benchmark on 33 Tasks: Adjusting Pessimism Per Suite



DoubleGum: simple, efficient, effective!

- Noise in Deep Q-Learning is shaped by two heteroscedastic Gumbel distributions
- Accounting for these distributions yields SOTA aggregate performance (AFAIK)
- Stable training across 33 continuous control environments

Questions: dythui2+drl@gmail.com
Code: https://github.com/dyth/doublegum

Thanks to Profs. Aaron Courville and Pierre-Luc Bacon + wider Mila community!