Principled Analysis of Deep Reinforcement Learning Evaluation and Design Paradigms

Ezgi Korkmaz

Abstract

Starting from the utilization of deep neural networks to approximate the state-action value function that led to winning one of the most challenging games, to algorithmic advancements that allowed solving problems without even explicitly stating the rules of the challenge at hand, reinforcement learning research has been the center of remarkable scientific progress for the past decade. In this paper, we focus on the key ingredients of this research progress and we analyze the canonical evaluation and design paradigms in reinforcement learning. We introduce the theoretical foundations of scaling laws in reinforcement learning and show that the asymptotic performance of reinforcement learning algorithms does not have a monotone relationship between performance rankings and data-regimes. We conduct large-scale experiments and our results demonstrate that a line of reinforcement learning research under the canonical design and evaluation paradigms resulted in incorrect conclusions. Our analysis and results provide a core analysis on scaling, capacity and complexity of deep reinforcement learning.

1 Introduction

Founded on rigorous theoretical guarantees, reinforcement learning research achieved high acceleration upon the proposal of the initial study on approximating the state-action value function via deep neural networks (Mnih et al. 2015; Stiennon et al. 2020; Schrittwieser et al. 2020; Lee et al. 2024; Korkmaz 2025). A line of highly successful deep reinforcement learning algorithms have been proposed (Hasselt, Guez, and Silver 2016; Wang et al. 2016; Hessel et al. 2018, 2021; Kapturowski et al. 2023; Korkmaz 2024) from focusing on different architectural ideas to foundations targeting overestimation, all of which were designed and tested in the high-data regime, i.e. two hundred million frame training. An alternative recent line of research with an extensive amount of publications focused on pushing the performance bounds of deep reinforcement learning policies in the low-data regime, i.e. with one hundred thousand environment interaction training. Many different concepts in current reinforcement learning research, from architectural proposals to learning underlying dynamics of the environment, experienced an accelerated progress and significant attention, growing into several major research fields, solely based on policy performance comparisons demonstrated in the low-data regime benchmark.

In this paper, we focus on evaluation paradigms, implicit assumptions and canonical methodological choices made in deep reinforcement learning research and demonstrate that there is a significant overlooked underlying premise driving this line of research without being explicitly discussed: that the performance profiles of deep reinforcement learning algorithms have a monotonic relationship with different sample-complexity regimes. We show that this implicit assumption, that is commonly shared amongst a large collection of low-data regime studies, shapes how the canonical design and evaluation choices are made in deep reinforcement learning research and represents a prominent misdirection in scientific progress. The suboptimal conclusions obtained from these canonical choices shape future research directions with incorrect reasoning. We show that these methodological decisions fuel incorrect justifications and conclusions, thereby misdirecting research efforts toward certain concepts for several years. Thus, in our paper we target these underlying premises and aim to answer the following questions:

What are the implicit assumptions and canonical choices in deep reinforcement learning research that fundamentally affect the conclusions made?

What is the foundational relationship between sample complexity and the algorithmic performance from the data-scarce regime to the asymptotic regime?

Hence, to be able to answer the questions raised above, in our paper we focus on underlying design and evaluation paradigms in deep reinforcement learning and make the following contributions:

- We analyze the evaluation paradigms and canonical methodological choices in deep reinforcement learning research, and introduce the theoretical foundations on how these methodological choices affect algorithm design, performance comparisons and algorithmic conclusions. Our analysis lays the foundations on scaling, capacity and complexity of deep reinforcement learning.
- Our theoretical analysis proves that the performance profile has a non-monotonic relationship with the asymptotic sample complexity and the low-data sample complexity regime. Regarding the central focus of the large scale implicit assumption instances, our results reveal that the

canonical methodological choices made in a line of deep reinforcement learning research have led to incorrect justifications and conclusions.

- We conduct large scale extensive experiments for a comprehensive and a diverse portfolio of deep reinforcement learning baseline algorithms in both the low-data regime and the high-data regime Arcade Learning Environment benchmark. Our results demonstrate that recent algorithms proposed and evaluated in the Arcade Learning Environment 100K benchmark are significantly affected by the implicit assumption on the relationship between performance profiles and sample complexity resulting in systematic bias in algorithmic evaluation.

2 Background and Preliminaries

The reinforcement learning problem is formalized as a Markov Decision Process (MDP) represented as a tuple $\langle S, A, \mathcal{P}, \mathcal{R}, \gamma, \rho_0 \rangle$ where S represents the state space, A represents the set of actions, $\mathcal{P}$ represents the transition probability distribution on $S \times A \times S$, $\mathcal{R} : S \times A \to \mathbb{R}$ represents the reward function, and $\gamma \in (0, 1]$ represents the discount factor. The aim in reinforcement learning is to learn an optimal policy $\pi(s, a)$ that maps state observations to actions $\pi : S \to \Delta(A)$, which maximizes the expected cumulative discounted rewards $R = \mathbb{E}_{a_t \sim \pi(s_t, \cdot)} \sum_t \gamma^t \mathcal{R}(s_t, a_t, s_{t+1})$. This objective is achieved by constructing a state-action value function that learns for each state-action pair the expected cumulative discounted rewards that will be obtained if action $a \in A$ is executed in state $s \in S$.

$$\mathcal{Q}(s, a) = \sum_{s'} \mathcal{P}(s'|s, a)[\mathcal{R}(s, a, s') + \gamma \mathcal{V}(s')]$$

In settings where the state space and/or action space is large enough that the state-action value function $\mathcal{Q}(s, a)$ cannot be held in a tabular form, a function approximator is used. Thus, for deep reinforcement learning the $\mathcal{Q}$-function is approximated via deep neural networks

$$\theta_{t+1} = \theta_t + \alpha(\mathcal{R}(s_t, a_t, s_{t+1}) + \gamma \mathcal{Q}(s_{t+1}, \arg\max_a \mathcal{Q}(s_{t+1}, a; \theta_t); \theta_t) - \mathcal{Q}(s_t, a_t; \theta_t)) \nabla_{\theta_t} \mathcal{Q}(s_t, a_t; \theta_t).$$

Dueling Architecture: The dueling architecture (Wang et al. 2016) outputs two streams of fully connected layers for both estimating the advantage $\mathcal{A}(s, a)$ for each action in a given state s, $\mathcal{A}(s, a) = \mathcal{Q}(s, a) - \max_a \mathcal{Q}(s, a)$ and the state values $\mathcal{V}(s)$. In particular, the last layer of the dueling architecture contains the forward mapping $\mathcal{Q}(s, a; \theta, \alpha, \beta) = \mathcal{V}(s; \theta, \beta) + \big(\mathcal{A}(s, a; \theta, \alpha) - \max_{a' \in A} \mathcal{A}(s, a'; \theta, \alpha)\big)$ where θ represents the parameters of the convolutional layers and α and β represent the parameters of the fully connected layers outputting the advantage and state value estimates respectively.

Inherent High-Capacity Models: The initial algorithm that has been proposed to have inherent high-capacity is C51. The projected Bellman update for the i^{th} atom is computed as

$$(\Phi \mathcal{T} \mathcal{Z}_\theta(s_t, a_t))_i = \sum_j^{\mathcal{N}-1} \big[1 - \frac{|[\mathcal{T} z_j]_{v_{\min}}^{v_{\max}} - z_i|}{\Delta z}\big]_0^1 \tau_j(s_{t+1}, \max_{a \in A} \mathbb{E} \mathcal{Z}_\theta(s_{t+1}, a))$$

where $z_i = v_{\min} + i\Delta z : 0 \leq i < \mathcal{N}$ represents the set of atoms in categorical learning, and the atom probabilities are learnt as a parametric model (Bellemare, Dabney, and Munos 2017)

$$\tau_i(s_t, \max_{a \in A} \mathbb{E} \mathcal{Z}_\theta(s_t, a)) = \frac{e^{\theta_i(s_t, a_t)}}{\sum_j e^{\theta_j(s_t, a_t)}}, \Delta z := \frac{v_{\max} - v_{\min}}{\mathcal{N} - 1}$$

Following this baseline the $\mathcal{Q}$RD$\mathcal{Q}$N algorithm (Dabney et al. 2018b) is proposed to learn the quantile projection

$$\mathcal{T} \mathcal{Z}(s_t, a_t) = \mathcal{R}(s_t, a_t, s_{t+1}) + \gamma \mathcal{Z}(s_{t+1}, \arg\max_{a \in A} \mathbb{E}_{z \sim \mathcal{Z}(s_{t+1}, a_{t+1})}[z])$$

with $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$ where $\mathcal{Z} \in Z$ represents the quantile distribution of an arbitrary value function. Following this study the I$\mathcal{Q}$N algorithm (Dabney et al. 2018a) is proposed to learn the full quantile function instead of learning a discrete set of quantiles as in the $\mathcal{Q}$RD$\mathcal{Q}$N algorithm. The I$\mathcal{Q}$N algorithm objective is to minimize the loss function

$$\mathcal{L} = \frac{1}{\mathcal{K}} \sum_{i=1}^{\mathcal{K}} \sum_{j=1}^{\mathcal{K}'} \rho_\delta(\mathcal{R}(s_t, a_t, s_{t+1}) + \gamma \mathcal{Z}_{\delta'_j}(s_{t+1}, \arg\max_{a \in A} \mathcal{Q}_\beta(s_t, a_t)) - \mathcal{Z}_{\delta_i}(s_t, a_t)) \quad (1)$$

where ρ_δ represents the Huber quantile regression loss, and $\mathcal{Q}_\beta = \int_0^1 \mathcal{F}_\mathcal{Z}^{-1}(\delta) d\beta(\delta)$. Note that $\mathcal{Z}_\delta = \mathcal{F}_\mathcal{Z}^{-1}(\delta)$ is the quantile function of the random variable $\mathcal{Z}$ at $\delta \in [0, 1]$.

3 Low-Data Regime Versus Asymptotic Performance

Our paper discovers both with extensive empirical analysis and theoretical investigation that asymptotic performance of reinforcement learning algorithms does not necessarily provide any information nor indication on their relative performance ranking in the low-data regime. The results provided in Section 6 extensively demonstrate that a large body of work in reinforcement learning research carried this assumption and resulted in incorrect conclusions. In this section, we introduce the foundational basis for our discovery revealed by our extensive empirical analysis in Section 6 in optimization of non-stationary policies, i.e. rewards and transitions that can vary with each step in an episode, in undiscounted, finite-horizon MDPs with linear function approximation. In particular, a finite horizon MDP is represented as a tuple $\langle S, A, \mathcal{P}, \mathcal{R}, \mathcal{H} \rangle$ where S is the set of states, and A represents the set of actions. For each time step $t \in [\mathcal{H}] = \{1, \ldots, \mathcal{H}\}$, state s, and action a the transition probability kernel $\mathcal{P}_t(s'|s, a)$ gives the probability distribution over the next state, and the reward $\mathcal{R}_t(s, a, s')$ gives the immediate rewards. A non-stationary policy $\pi = (\pi_1, \ldots, \pi_\mathcal{H})$ induces a state-action value function given by

$$\mathcal{Q}_t^\pi(s, a) = \mathbb{E}\left[\sum_{h=t}^{\mathcal{H}} \mathcal{R}_h(s_h, \pi_h(s_h), s_{h+1}) \middle| s_h = s, a_h = a\right]$$

where we let $a_h \sim \pi_h(s_h)$, and the corresponding value function $\mathcal{V}_t^\pi(s) = \mathcal{Q}_t(s, \pi_t(s))$. The optimal non-stationary policy π^* has value function $\mathcal{V}_t^*(s) = \mathcal{V}_t^{\pi^*}(s)$ satisfying $\mathcal{V}_t^*(s) = \sup_\pi \mathcal{V}_t^\pi(s)$. The objective is to learn a sequence of non-stationary policies π^k for $k \in \{1, \ldots, \mathcal{K}\}$ while interacting with an unknown MDP in order to minimize the

regret, which is measured asymptotically over $\mathcal{K}$ episodes of length $\mathcal{H}$, $\text{REGRET}(\mathcal{K}) = \sum_{k=1}^{\mathcal{K}} \left(\mathcal{V}_1^*(s_1^k) - \mathcal{V}_1^{\pi^k}(s_1^k)\right)$ where $s_1^k \in S$ is the starting state of the k-th episode. Regret sums up the gap between the expected rewards obtained by the sequence of learned policies π^k and those obtained by π^* when learning for $\mathcal{K}$ episodes. In the linear function approximation setting there is a feature map $\phi_t : S \times A \to \mathbb{R}^{d_t}$ for each $t \in [\mathcal{H}]$ that sends a state-action pair (s, a) to the d_t-dimensional vector $\phi_t(s, a)$. Then, the state-action value function $\mathcal{Q}_t(s, a)$ is parameterized by a vector $\theta_t \in \mathbb{R}^{d_t}$ so that $\mathcal{Q}_t(\theta_t)(s, a) = \phi_t(s, a)^\top \theta_t$. Recent theoretical work in this setting gives an algorithm along with a lower bound that matches the regret achieved by the algorithm up to logarithmic factors.

Theorem 3.1 ((Zanette et al. 2020))**.** *Under appropriate normalization assumptions there is an algorithm that learns a sequence of policies π^k achieving regret $\text{REGRET}(\mathcal{K}) = \tilde{O}\left(\sum_{t=1}^{\mathcal{H}} d_t\sqrt{\mathcal{K}} + \sum_{t=1}^{\mathcal{H}} \sqrt{d_t}\mathcal{I}\mathcal{K}\right)$, where $\mathcal{I}$ is the inherent Bellman error. Furthermore, this regret bound is optimal for this setting up to logarithmic factors in $d_t, \mathcal{K}$ and $\mathcal{H}$ whenever $\mathcal{K} = \Omega((\sum_{t=1}^{\mathcal{H}} d_t)^2)$, in the sense that for any level of inherent Bellman error $\mathcal{I}$ and sequence of feature dimensions $\{d_t\}_{t=1}^H$, there exists a class of MDPs $\mathcal{C}(\mathcal{I}, \{d_t\}_{t=1}^H)$ where any algorithm achieves at least as much regret on at least one MDP in the class.*

The class of MDPs $\mathcal{C}(\mathcal{I}, \{d_t\}_{t=1}^H)$ constructed in Theorem 3.1 additionally satisfies the following properties. First, every MDP in $\cup_{\mathcal{I}, \{d_t\}_{t=1}^H} \mathcal{C}(\mathcal{I}, \{d_t\}_{t=1}^H)$ has the same transitions (up to renaming of states and actions). Second, for each fixed value of the inherent Bellman error $\mathcal{I}$ and the dimensions $\{d_t\}_{t=1}^H$, every MDP in $\mathcal{C}(\mathcal{I}, \{d_t\}_{t=1}^H)$ utilizes the same feature map $\phi_t(s_t, a_t)$. Thus one can view the class $\mathcal{C}(\mathcal{I}, \{d_t\}_{t=1}^H)$ as encoding one "underlying" true environment defined by the transitions, with varying values of $\mathcal{I}$ and $\{d_t\}_{t=1}^H$ corresponding to varying levels of function approximation accuracy, and model capacity for the underlying environment. For simplicity of notation we will focus on the setting where $d_t = d$ for all $t \in \{1, \dots H\}$ and write $\mathcal{C}(\mathcal{I}, d)$ for the class of MDPs constructed in Theorem 3.1 for this setting. Utilizing this point of view, we can then prove the following theorem on the relationship between the performance in the asymptotic and low-data regimes.

Theorem 3.2 (*Non-monotonocity Across Regimes*)**.** *For any $\epsilon > 0$, let d_α be any feature dimension, and let $d_\beta = d_\alpha^{1-\epsilon/2}$. Then there exist thresholds $\mathcal{K}_{low} < \mathcal{K}_{high}$ and inherent Bellman error levels $\mathcal{I}_\beta > \mathcal{I}_\alpha$ such that*

1. *There is an algorithm achieving regret $\text{REGRET}_{low}(\mathcal{K})$ when $\mathcal{K} < \mathcal{K}_{low}$ for all MDPs in $\mathcal{C}(\mathcal{I}_\beta, d_\beta)$. However, every algorithm has regret at least $\widetilde{\Omega}\left(d_\beta^{\epsilon/2}\text{REGRET}_{low}(\mathcal{K})\right)$ when $\mathcal{K} < \mathcal{K}_{low}$ on some MDP $M \in \mathcal{C}(\mathcal{I}_\alpha, d_\alpha)$.*
2. *There is an algorithm achieving regret $\text{REGRET}_{high}(\mathcal{K})$ when $\mathcal{K} > \mathcal{K}_{high}$ for all MDPs in $\mathcal{C}(\mathcal{I}_\alpha, d_\alpha)$. However, every algorithm has regret at least $\widetilde{\Omega}\left(d_\alpha^{\epsilon}\text{REGRET}_{high}(\mathcal{K})\right)$ on some MDP $M \in \mathcal{C}(\mathcal{I}_\beta, d_\beta)$ when $\mathcal{K} > \mathcal{K}_{high}$.*

Proof. Let $\epsilon > 0$ and consider $d_\beta = d_\alpha^{1-\frac{\epsilon}{2}}, \mathcal{I}_\beta = \frac{1}{d_\alpha^\epsilon\sqrt{d_\beta}}, \mathcal{I}_\alpha = \frac{1}{d_\alpha^{\frac{1}{2}+2\epsilon}}, \mathcal{K}_{\text{low}} = d_\alpha^{2+\epsilon}, \mathcal{K}_{\text{high}} = d_\alpha^{2+4\epsilon}$ We begin with the proof of part 1. Therefore, for $\mathcal{K} < \mathcal{K}_{\text{low}}$, $\sqrt{d_\beta}\mathcal{I}_\beta\mathcal{K} = d_\alpha^{-\epsilon}\mathcal{K} < d_\alpha^{1-\frac{\epsilon}{2}}\sqrt{\mathcal{K}} = d_\beta\sqrt{\mathcal{K}}$. Therefore, by Theorem 3.1 there exists an algorithm achieving regret

$$\begin{aligned}\text{REGRET}_{\text{low}}(\mathcal{K}) &= \tilde{O}\left(\mathcal{H}d_\beta\sqrt{\mathcal{K}} + \mathcal{H}\sqrt{d_\beta}\mathcal{I}_\beta\mathcal{K}\right)\\ &= \widetilde{O}\left(d_\beta\sqrt{\mathcal{K}}\right)\end{aligned}$$

in every MDP $M \in \mathcal{C}(\mathcal{I}_\beta, d_\beta)$. Further, since $\mathcal{K}_{\text{low}} = d_\alpha^{2+\epsilon} > \widetilde{\Omega}\left(d_\alpha^2\right)$, the lower bound from Theorem 3.1 applies to the class of MDPs $\mathcal{C}(\mathcal{I}_\alpha, d_\alpha)$ for all $\mathcal{K} \in \left[\widetilde{\Omega}\left(d_\alpha^2\right), \mathcal{K}_{\text{low}}\right]$. In particular, every algorithm receives regret at least

$$\begin{aligned}\text{REGRET}(\mathcal{K}) &= \widetilde{\Omega}\left(\mathcal{H}d_\alpha\sqrt{\mathcal{K}} + \mathcal{H}\sqrt{d_\alpha}\mathcal{I}_\alpha\mathcal{K}\right)\\ &> \widetilde{\Omega}\left(\mathcal{H}d_\beta^{\frac{1}{1-\epsilon/2}}\sqrt{\mathcal{K}}\right) > \widetilde{\Omega}\left(\mathcal{H}d_\beta^{\frac{\epsilon/2}{1-\epsilon/2}}d_\beta\sqrt{\mathcal{K}}\right)\end{aligned}$$

Thus, $\text{REGRET}(\mathcal{K}) > \widetilde{\Omega}\left(d_\beta^{\epsilon/2}\text{REGRET}_{\text{low}}(\mathcal{K})\right)$. For part 2, note that for $\mathcal{K} > \mathcal{K}_{\text{high}}$ we have both $\sqrt{d_\alpha}\mathcal{I}_\alpha\mathcal{K} = d_\alpha^{-2\epsilon}\mathcal{K} > d_\alpha^{-2\epsilon}\sqrt{\mathcal{K}\cdot\mathcal{K}_{\text{high}}} > d_\alpha\sqrt{\mathcal{K}}$ and $\sqrt{d_\beta}\mathcal{I}_\beta\mathcal{K} > d_\alpha^{-\epsilon}\sqrt{\mathcal{K}\cdot\mathcal{K}_{\text{low}}} = d_\alpha^{1+\epsilon}\sqrt{\mathcal{K}} > d_\beta\sqrt{\mathcal{K}}$. Therefore by Theorem 3.1 that for $\mathcal{K} > \mathcal{K}_{\text{high}}$ there exists an algorithm achieving regret

$$\begin{aligned}\text{REGRET}_{\text{high}}(\mathcal{K}) &= \tilde{O}\left(\mathcal{H}d_\alpha\sqrt{\mathcal{K}} + \mathcal{H}\sqrt{d_\alpha}\mathcal{I}_\alpha\mathcal{K}\right)\\ &= \tilde{O}\left(\mathcal{H}\sqrt{d_\alpha}\mathcal{I}_\alpha\mathcal{K}\right).\end{aligned}$$

for every MDP $M \in \mathcal{C}(\mathcal{I}_\alpha, d_\alpha)$. However, by the lower bound in Theorem 3.1, for $\mathcal{K} > \mathcal{K}_{\text{high}}$ every algorithm receives regret at least

$$\begin{aligned}\text{REGRET}(\mathcal{K}) &= \widetilde{\Omega}\left(\mathcal{H}d_\beta\sqrt{\mathcal{K}} + \mathcal{H}\sqrt{d_\beta}\mathcal{I}_\beta\mathcal{K}\right)\\ &> \widetilde{\Omega}\left(\mathcal{H}\sqrt{d_\beta}\mathcal{I}_\beta\mathcal{K}\right) = \widetilde{\Omega}\left(\mathcal{H}d_\alpha^{-\epsilon}\mathcal{K}\right)\\ &= \widetilde{\Omega}\left(d_\alpha^{\epsilon}\mathcal{H}d_\alpha^{-2\epsilon}\mathcal{K}\right) = \widetilde{\Omega}\left(d_\alpha^{\epsilon}\mathcal{H}\sqrt{d_\alpha}\mathcal{I}_\alpha\mathcal{K}\right)\\ &> \widetilde{\Omega}\left(d_\alpha^{\epsilon}\text{REGRET}_{\text{high}}(\mathcal{K})\right)\end{aligned}$$

□

Theorem 3.2 introduces the provable trade-off between performance in the low-data regime, i.e. $\mathcal{K} < \mathcal{K}_{\text{low}}$, and the high-data regime, i.e. $\mathcal{K} > \mathcal{K}_{\text{high}}$. In particular, in the low-data regime lower capacity function approximation, i.e. lower feature dimension d_β, with larger approximation error, i.e. larger inherent Bellman error $\mathcal{I}_\beta$, can provably outperform larger capacity models, i.e. feature dimension d_α, with smaller approximation error, i.e. inherent Bellman error $\mathcal{I}_\alpha$. Furthermore, the relative performance is reversed in the high-data regime $\mathcal{K} > \mathcal{K}_{\text{high}}$. Thus, asymptotic performance of an algorithm is neither indicative nor carries any relevant information on the expected performance of the algorithm when training data is scarce (i.e. limited).

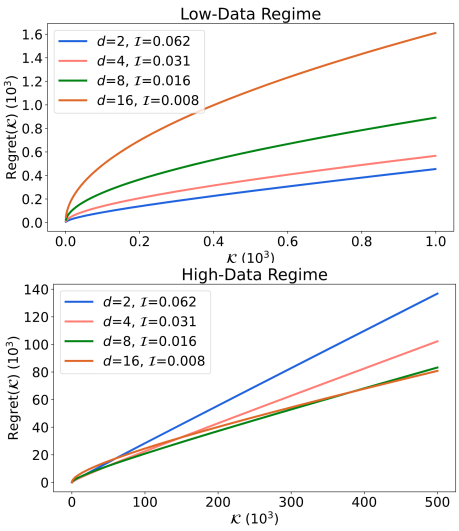
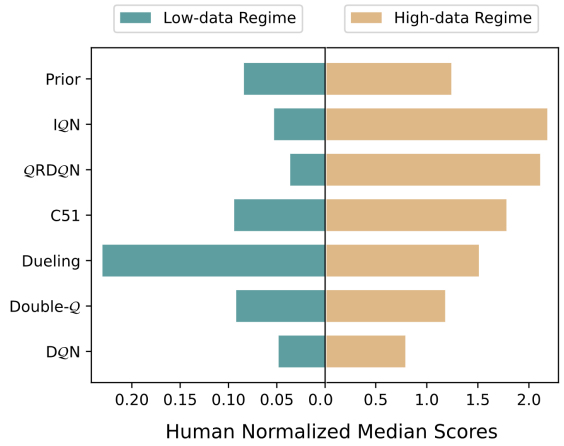


Figure 1: Up: Regret in the low-data regime. Down: Regret in the high-data regime.



Asymptotic vs Low-data

Figure 2: Scaling laws of reinforcement learning: Baseline comparison of algorithms that were proposed and developed in the high-data regime in the Arcade Learning Environment in both high-data regime and low-data regime.

4 The Assumption of Monotonicity and Performance Rankings

The instances of the implicit assumption that the performance profile of an algorithm in the high-data regime will translate to the low-data regime monotonically appear in almost all of the studies conducted in the low-data regime. In particular, we see that when this line of work was being conducted the best performing algorithm in the high-data regime was an inherently high capacity model, i.e. based on learning the state action value distribution. Hence, there are many cases in the literature (e.g. DRQ, OTR, DER, CURL, SimPLE, Efficient-Zero) where all the newly proposed algorithms in the low-data regime are being compared to an algorithm that inherently produces a higher capacity model **under the implicit assumption** that an algorithm that is state-of-the art in the high-data regime must be the state-of-the-art in the low-data regime. The large scale experiments provided in Section 6 demonstrate the impact of this implicit assumption and provide a guideline for a principled analysis and evaluation. In particular, the results reported in Section 6 prove that the performance profile of an algorithm in the high-data regime does not monotonically transfer to the low-data regime. Due to this extensive focus throughout the literature on low-data regime comparisons to algorithms that inherently learn higher capacity models, we provide additional theoretical analysis for the empirically observed sample complexity results in the low to high-data regime in deep reinforcement learning. The following proposition demonstrates a precise justification of these issues: whenever there are two different actions where the true mean state-action values are within ϵ, an approximation error of ϵ in total variation distance d_{TV} for $\mathcal{D}(s,a)$ of one of the actions can be sufficient to reverse the order of the means.

Proposition 4.1 (*Sufficiency of error of ϵ in total variation distance*)**.** *Fix a state s and consider two actions $a, \hat{a}$. Let $\mathcal{D}(s,a)$ be the true state-action value distribution of (s,a), and let $\mathcal{Z}(s,a) \sim \mathcal{D}(s,a)$. Suppose that $\mathbb{E}[\mathcal{Z}(s,a)] = \mathbb{E}[\mathcal{Z}(s,\hat{a})] + \epsilon$. Then there is a random variable $\mathcal{Y}$ such that $d_{TV}(\mathcal{Y}, \mathcal{Z}(s,a)) \leq \epsilon$ and $\mathbb{E}[\mathcal{Z}(s,\hat{a})] \geq \mathbb{E}[\mathcal{Y}]$.*

The proof is provided in the supplementary material. Proposition 4.1 shows that to have the correct ranking of the actions the state-action value distribution must be learnt with error at most ϵ. Standard results on sample complexity for discrete distributions then imply that algorithms that learn the state-action value distribution with fixed support size k, i.e C51, require k/ϵ^2 samples to achieve total variation distance at most ϵ. More advanced algorithms such as $\mathcal{Q}$RD$\mathcal{Q}$N and I$\mathcal{Q}$N do away with the assumption that the support is known. This allows a more flexible representation in order to more accurately represent state-action values, but, as we will show, leads to a further increase in the sample complexity. The $\mathcal{Q}$RD$\mathcal{Q}$N algorithm models it as a uniform mixture of $\mathcal{N}$ Dirac deltas on the reals i.e. $\mathcal{Z}(s,a) = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i(s,a)}$, where $\theta_i(s,a) \in \mathbb{R}$ is a parametric model.

Proposition 4.2 (*Sample Complexty with Unknown Support*)**.** *Let $\mathcal{N} > \mathcal{M} \geq 2$, $\epsilon > \frac{\mathcal{M}}{4\mathcal{N}}$, and $\theta_i \in \mathbb{R}$ for $i \in [\mathcal{N}]$. The number of samples required to learn a model of the form $\mathcal{Z} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} \delta_{\theta_i}$ to within total variation distance ϵ is $\Omega\left(\frac{\mathcal{M}}{\epsilon^2}\right)$.*

The proof is provided in the supplementary material. Note that the lower bound in Proposition 4.2 can be significantly larger than k/ϵ^2 samples.

5 Principled Evaluation Framework

In Section 6, we systematically explain and discuss the underlying design paradigms, the implicit assumptions and the methodological choices made in deep reinforcement learning research that led to incorrect conclusions. In this section we introduce the principled evaluation framework to ensure the research progress we obtain in deep reinforcement learning is reliable and scientifically robust.

I. Assumptions matter: Performance rankings across regimes are non-monotone.
II. Biases in Evaluation: Including algorithms in the comparison benchmark based on the monotonicity assumption will create biased evaluation.
III. Core Algorithms: Core algorithms must be included in the comparison benchmarks.
IV. Inherent Capacity: Inherent capacity and dimensionality will provide insights on performance rankings across regimes.
V. Biases in Datasets: Creating datasets based on the monotonicity assumption will create biased benchmarks.

6 Large Scale Empirical Analysis

The empirical analysis is conducted in the Arcade Learning Environment (ALE) (Mnih et al. 2015). The Double $\mathcal{Q}$-learning algorithm is trained via Hasselt, Guez, and Silver (2016) initially proposed by van Hasselt (2010). The dueling algorithm is trained via Wang et al. (2016). The prior algorithm refers to the prioritized experience replay algorithm proposed by Schaul et al. (2016). The experiments are run with Haiku as the neural network library, Optax (Hessel et al. 2020) as the optimization library, and RLax for the reinforcement learning library (Babuschkin et al. 2020). All of the results are reported with the standard error of the mean. For the full list of algorithms, details on the hyperparameters, direct references and the detailed explanations of the baselines please see the supplementary material. To provide a complete picture of the sample complexity we conducted our experiments in both low-data, i.e. the Arcade Learning Environment 100K benchmark, and high data regime, i.e. baseline 200 million frame training. Note that human normalized score is computed as follows: $\text{Score}_{\text{HN}} = (\text{Score}_{agent} - \text{Score}_{random})/(\text{Score}_{human} - \text{Score}_{random})$.

Implicit Assumptions on Monotonicity Cause Suboptimal and Incorrect Conclusions. Our extensive large-scale empirical analysis demonstrates that a major line of research conducted in the past five years resulted in incorrect conclusions. We show that a simple baseline algorithm **from 2016** (Wang et al. 2016), by a systematic methodological choice was never included in the comparison benchmark, following the implicit assumption that appears in all of the recent line of research that we have discussed in detail in Section 4. We demonstrate that this baseline algorithm in fact performs much better than many recent algorithms that claimed to be better than the baselines, even including algorithms that are specifically built on top of the baseline algorithm. Figure 4 reports learning curves for the I$\mathcal{Q}$N, $\mathcal{Q}$RD$\mathcal{Q}$N, dueling and C51 in the Arcade Learning Environment low-data regime benchmark. These results demonstrate that the simple base algorithm dueling performs significantly better than a series of algorithms that were included in the comparison benchmark which inherently produced higher capacity models when the training samples are limited. Note that DR$\mathcal{Q}$ uses the dueling architecture without any high capacity inducing components. One intriguing takeaway from the results provided in Table 1 and the results reported in the full version of our paper [1] is the fact that the simple baseline dueling algorithm performs 15% better than the DR$\mathcal{Q}^{\text{NeurIPS}}$ implementation, and 11% less than the DR$\mathcal{Q}^{\text{ICLR}}$ implementation instead of 82% gain reported in the original paper.

Providing Direct Comparison to Core Algorithms. Algorithms that are built on top of a core reinforcement learning algorithm must provide a direct comparison to the algorithm they are built on top of. The case of DR$\mathcal{Q}$ demonstrates the significance of the direct comparison to the core algorithm. As our paper discovers and describes extensively, the monotonicity assumption on the performance ranking across regimes led a line of work to benchmark against certain algorithms in the low-data regime, assuming that if an algorithm has the highest performance in the high-data regime it must have the top-ranked performance in the low-data regime. However, as we pointed out in our theoretical analysis this is a dangerous and incorrect assumption. The results reported in Figure 2 demonstrate that these implicit assumptions in fact lead to incorrect and suboptimal conclusions.

Non-Monotonicity of Performance Ranking Across Regimes. Table 1 reports the human normalized median, mean and 20$^{\text{th}}$ percentile results over all of the MDPs from the 100K ALE benchmark for D$\mathcal{Q}$N, Double-$\mathcal{Q}$, dueling, C51, $\mathcal{Q}$RD$\mathcal{Q}$N, I$\mathcal{Q}$N and Prior. One important takeaway from the results reported in the Table 1 is the fact that one particular algorithm performance profile in 200 million frame training will not directly transfer to the low-data region as predicted by our theoretical analysis in Section 3. Figure 3 reports the learning curves of human normalized median, mean and 20$^{\text{th}}$ percentile for the dueling algorithm, C51, $\mathcal{Q}$RD$\mathcal{Q}$N, and I$\mathcal{Q}$N. These results once more demonstrate that the performance profile of the simple base algorithm dueling is significantly better than any core algorithm which inherently produced higher capacity models that was included in the comparison benchmark of the extensive low-data regime literature when the number of environment interactions are limited.

Biases in the Evaluation Criteria. The original paper of the DR$\mathcal{Q}^{\text{ICLR}}$ algorithm (Yarats, Kostrikov, and Fergus 2021) benchmarks against data-efficient Rainbow (DER) (van Hasselt, Hessel, and Aslanides 2019) which inherently learns a higher capacity model. Our results show that the fact that the original paper that proposed data augmentation for reinforcement learning, i.e. DR$\mathcal{Q}^{\text{ICLR}}$, on top of the dueling algorithm did not provide comparisons against the core algorithm that they are built on, i.e. dueling (Wang et al. 2016), resulted in inflated performance profiles for the DR$\mathcal{Q}^{\text{ICLR}}$ algorithm. For a fair, direct and transparent comparison we kept the hyperparameters for the baseline algorithms in the low-data regime exactly the same with the DR$\mathcal{Q}^{\text{ICLR}}$ paper (see supplementary material for the full list and high-data regime hyperparameter settings). More intriguingly, the comparisons provided in the

[1] DER^{2021} refers to the re-implementation with random seed variations of the original paper data-efficient Rainbow (i.e. DER^{2019}) by van Hasselt, Hessel, and Aslanides (2019). OTR refers to further implementation of the Rainbow algorithm by Kielak (2019). DR$\mathcal{Q}^{\text{NeurIPS}}$ refers to the re-implementation of the original DR$\mathcal{Q}$ algorithm with the goal of achieving reproducibility with variation on the number of random seeds (Agarwal et al. 2021).

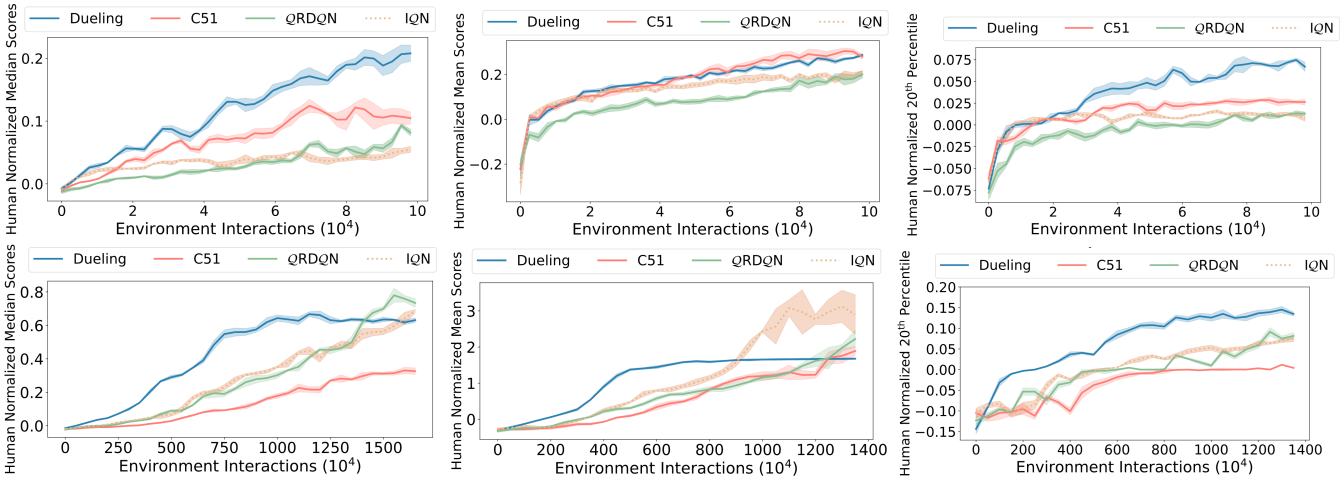


Figure 3: Up: Human normalized median, mean and 20$^{\text{th}}$ percentile results for the dueling algorithm, C51, I$\mathcal{Q}$N and $\mathcal{Q}$RD$\mathcal{Q}$N in the Arcade Learning Environment 100K benchmark. Down: Human normalized median, mean, and 20$^{\text{th}}$ percentile results for the dueling algorithm, C51, I$\mathcal{Q}$N and $\mathcal{Q}$RD$\mathcal{Q}$N in the high-data regime towards 200 million frame.

Algorithms	Human Normalized Median	Human Normalized Mean	20$^{\text{th}}$ Percentile
D$\mathcal{Q}$N	0.0481±0.0036	0.1535±0.0119	0.0031±0.0032
Double-$\mathcal{Q}$	0.0920±0.0181	**0.3169±0.0196**	0.0341±0.0042
Dueling	**0.2304±0.0061**	0.2923±0.0060	**0.0764±0.0037**
C51	0.0941±0.0081	0.3106±0.0199	0.0274±0.0024
$\mathcal{Q}$RD$\mathcal{Q}$N	0.0820±0.0037	0.2171±0.0098	0.0189±0.0031
I$\mathcal{Q}$N	0.0528±0.0058	0.2050±0.0123	0.0091±0.0011
Prior	0.0840±0.0018	0.2792±0.0123	0.0267±0.0042

Table 1: Large scale comparison of $\mathcal{Q}$-based deep reinforcement learning algorithms with human normalized mean, median and 20$^{\text{th}}$ percentile results in the Arcade Learning Environment 100K benchmark for D$\mathcal{Q}$N (Mnih et al. 2015), deep Double-$\mathcal{Q}$ (Hasselt, Guez, and Silver 2016), dueling (Wang et al. 2016), Prior (Schaul et al. 2016), C51, $\mathcal{Q}$RD$\mathcal{Q}$N and I$\mathcal{Q}$N (Dabney et al. 2018a).

DR$\mathcal{Q}^{\text{ICLR}}$ paper to the DER and OTR algorithms report the performance gained by DR$\mathcal{Q}^{\text{ICLR}}$ over DER is 82% and over OTR is 35%. However, if a direct comparison is made to the simple dueling algorithm as Table 1 demonstrates the performance gain is utterly restricted to **11%**. Moreover, when it is compared to the reproduced results of DR$\mathcal{Q}^{\text{NeurIPS}}$ our results reveal that in fact there is a performance decrease due to utilizing DR$\mathcal{Q}$ over dueling. Thus, while our paper introduces the foundations on the non-monotonicity of the performance profiles from large-data regime to low-data regime, it further provides the basis on how we can compare algorithms with a principled approach and scientific rigor allowing more concrete and accurate evaluation across data-regimes.

Theoretical Analysis and the Inherent Bellman Error vs Dimensionality. The right and center plots of Figure 2 report regret curves corresponding to the theoretical analysis in Theorem 3.2 for various choices of the feature dimensionality d and the inherent Bellman error $\mathcal{I}$. In particular, the center plot shows the low-data regime where the number of episodes $K < 1000$, while the right plot shows the high-data regime where K is as large as 500000. Notably, the relative ordering of the regret across the different choices of d and $\mathcal{I}$ is completely reversed in the high-data regime when compared to the low-data regime. Recall from Theorem 3.1 that the inherent Bellman error is a measure of the accuracy of function approximation under the Bellman operator corresponding to an MDP. Thus, the varying values of $\mathcal{I}$ and d in Figure 2 correspond to a natural setting where increasing the number of model parameters (i.e. increasing d) corresponds to an increase in the accuracy of function approximation (i.e. a decrease in $\mathcal{I}$). Thus the results reported in Figure 2 demonstrate that, even in the natural setting where increased model capacity leads to increased accuracy, there can be a complete reversal in the ordering of algorithm performance between the low and high-data regimes. The full version of our paper reports results on the number of samples required for training with the baseline algorithm that inherently produces higher capacity models to reach the same performance levels achieved by the dueling algorithm for every MDP from ALE low-data regime benchmark. These results once more demonstrate that to reach the same performance levels with the dueling algorithm, baseline algorithms that inherently learn higher capacity models require orders of magnitude more samples to train on. As discussed in Section 4, more complex representations for broader classes of distributions come at the cost of a higher sample complexity required for learning. One intriguing fact is that the original SimPLE paper in

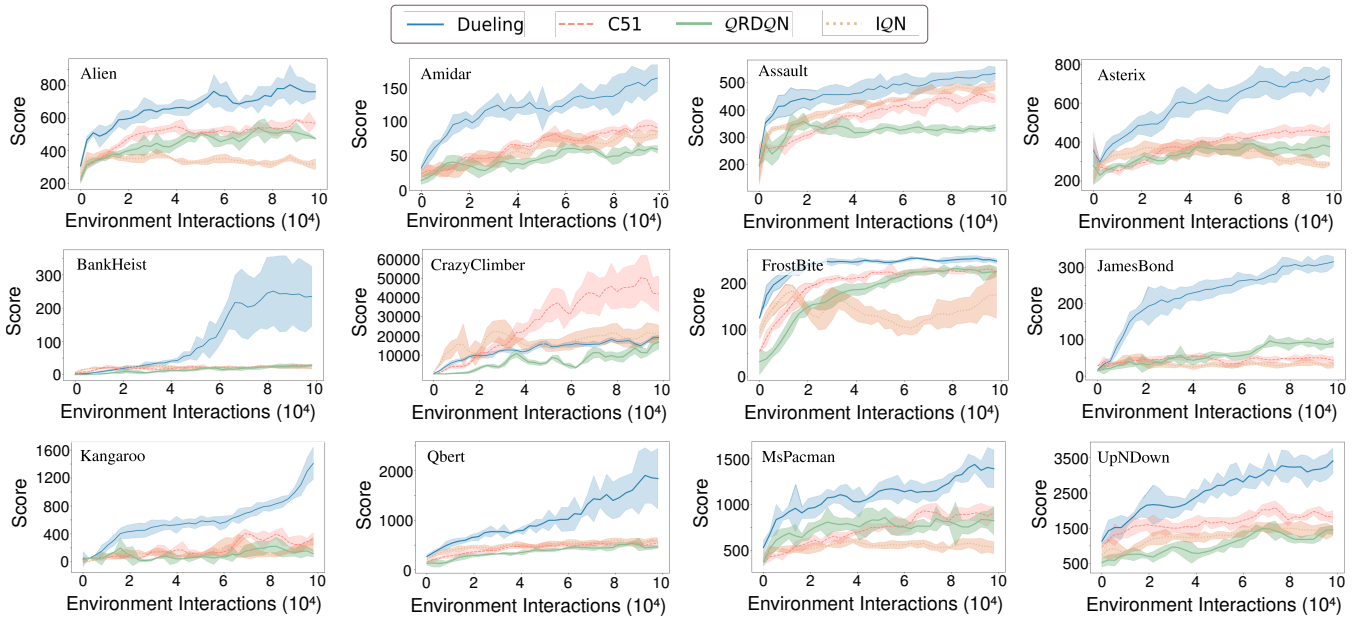


Figure 4: The learning curves of Alien, Amidar, Asterix, BankHeist, CrazyClimber, JamesBond, Kangaroo, MsPacman, FrostBite, Qbert, RoadRunner, and UpNDown with dueling architecture, C51, IQN and QRDQN algorithms in the Arcade Learning Environment with 100K environment interaction training.

the low-data regime benchmarked against the Rainbow algorithm which is essentially a higher capacity model designed in the high-data regime by having the implicit assumption that the state-of-the art performance profile must transfer monotonically to the low-data regime. These instances of implicit assumptions also occur in DR$\mathcal{Q}^{\text{ICLR}}$, CURL, SPR and Efficient-Zero (Ye et al. 2021) even when comparisons are made for more advanced algorithms such as MuZero.

Datasets are Created and Founded on Implicit Assumptions. Thus far we have discussed the pivotal role of implicit assumptions on the algorithmic comparisons and developing baselines in deep reinforcement learning. However, this issue further extends back to even how the entire low-data regime benchmark was established, i.e. ALE 100K. The ALE 100K was initially created to allow researchers to work on a subset of games instead of full set of games used in the high-data regime (Kaiser et al. 2020), and this benchmark is currently used by any algorithm developed for the low data regime. However, the entire ALE 100K benchmark was in fact built on the selection bias of choosing games that performed better either with the proposed algorithm of the paper that proposed the entire benchmark (Kaiser et al. 2020), or with Rainbow, which we extensively demonstrated throughout the paper is an algorithm that is subjected to the implicit assumption bias on monotonicity across regimes. Thus the issues we explicitly discover and analyze in our paper are not limited to baselines but further extend to canonical benchmarks that we evaluate reinforcement learning algorithms on. Our paper discovers that the canonical methodological choices made in a major line of deep reinforcement learning research that is based on these implicit assumptions, give incorrect signals on why and what makes these algorithms work, and hence affect future research directions while misdirecting the possible current research efforts from ideas that could have worked during the algorithm design process.

7 Conclusion

In this paper we aimed to answer the following questions: *(i) How are the scaling laws of reinforcement learning formally characterized with respect to capacity and complexity? (ii) What are the canonical methodological choices that fundamentally affect the progress in deep reinforcement learning research, (iii) What is the underlying theoretical relationship between monotonicity, the performance profiles and sample complexity regimes?* To be able to answer these questions we provide theoretical analysis on the sample complexity of the baseline deep reinforcement learning algorithms. We conduct extensive experiments both in the low-data regime 100K Arcade Learning Environment and high-data regime baseline 200 million frame training. Both theoretical and empirical analysis provided in our paper demonstrate that the performance profiles of deep reinforcement learning algorithms do not have a monotonic relationship across sample complexity regimes. Our analysis reveals that the underlying assumption of the monotonic relationship of the performance characteristics and the sample complexity regimes is currently present in a major line of research including many recent state-of-the-art studies and this implicit assumption led these studies to result in incorrect conclusions. Our results demonstrates that several baseline $\mathcal{Q}$ algorithms perform better than a line of recent algorithms claimed to be the state-of-the-art. Our paper establishes a principled analysis of deep reinforcement learning that characterizes the fundamental relationship between scaling, capacity and complexity.

References

Agarwal, R.; Schwarzer, M.; Castro, P. S.; Courville, A. C.; and Bellemare, M. G. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, 29304–29320.

Babuschkin, I.; Baumli, K.; Bell, A.; Bhupatiraju, S.; Bruce, J.; Buchlovsky, P.; Budden, D.; Cai, T.; Clark, A.; Danihelka, I.; Fantacci, C.; Godwin, J.; Jones, C.; Hennigan, T.; Hessel, M.; Kapturowski, S.; Keck, T.; Kemaev, I.; King, M.; Martens, L.; Merzic, H.; Mikulik, V.; Norman, T.; Quan, J.; Papamakarios, G.; Ring, R.; Ruiz, F.; Sanchez, A.; Schneider, R.; Sezener, E.; Spencer, S.; Srinivasan, S.; Stokowiec, W.; and Viola, F. 2020. The DeepMind Ecosystem.

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A Distributional Perspective on Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 449–458. PMLR.

Dabney, W.; Ostrovski, G.; Silver, D.; and Munos, R. 2018a. Implicit Quantile Networks for Distributional Reinforcement Learning. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 1104–1113. PMLR.

Dabney, W.; Rowland, M.; Bellemare, M. G.; and Munos, R. 2018b. Distributional Reinforcement Learning With Quantile Regression. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2892–2901. AAAI Press.

Hasselt, H. v.; Guez, A.; and Silver, D. 2016. Deep Reinforcement Learning with Double Q-Learning. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Hessel, M.; Budden, D.; Viola, F.; Rosca, M.; Sezener, E.; and Hennigan, T. 2020. Optax: composable gradient transformation and optimisation.

Hessel, M.; Danihelka, I.; Viola, F.; Guez, A.; Schmitt, S.; Sifre, L.; Weber, T.; Silver, D.; and van Hasselt, H. 2021. Muesli: Combining Improvements in Policy Optimization. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 4214–4226. PMLR.

Hessel, M.; Modayil, J.; van Hasselt, H.; Schaul, T.; Ostrovski, G.; Dabney, W.; Horgan, D.; Piot, B.; Azar, M. G.; and Silver, D. 2018. Rainbow: Combining Improvements in Deep Reinforcement Learning. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 3215–3222. AAAI Press.

Kaiser, L.; Babaeizadeh, M.; Milos, P.; Osinski, B.; Campbell, R. H.; Czechowski, K.; Erhan, D.; Finn, C.; Kozakowski, P.; Levine, S.; Mohiuddin, A.; Sepassi, R.; Tucker, G.; and Michalewski, H. 2020. Model Based Reinforcement Learning for Atari. In *8th International Conference on Learning Representations, ICLR 2020 [Spotlight Presentation]*.

Kapturowski, S.; Campos, V.; Jiang, R.; Rakicevic, N.; van Hasselt, H.; Blundell, C.; and Badia, A. P. 2023. Human-level Atari 200x faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Kielak, K. P. 2019. Do recent advancements in model-based deep reinforcement learning really improve data efficiency? *CoRR*.

Korkmaz, E. 2024. Understanding and Diagnosing Deep Reinforcement Learning. In *International Conference on Machine Learning, ICML 2024*.

Korkmaz, E. 2025. Counteractive RL: Rethinking Core Principles for Efficient and Scalable Deep Reinforcement Learning. *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025 [Spotlight Presentation]*.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2024. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, a. G.; Graves, A.; Riedmiller, M.; Fidjeland, A.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518: 529–533.

Schaul, T.; Quan, J.; Antonogloua, I.; and Silver, D. 2016. Prioritized Experience Replay. *International Conference on Learning Representations (ICLR)*.

Schrittwieser, J.; Antonoglou, I.; Hubert, T.; Simonyan, K.; Sifre, L.; Schmitt, S.; Guez, A.; Lockhart, E.; Hassabis, D.; Graepel, T.; Lillicrap, T. P.; and Silver, D. 2020. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nat.*, 588(7839): 604–609.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

van Hasselt, H. 2010. Double Q-learning. In Lafferty, J. D.; Williams, C. K. I.; Shawe-Taylor, J.; Zemel, R. S.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information*

Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada, 2613–2621. Curran Associates, Inc.

van Hasselt, H.; Hessel, M.; and Aslanides, J. 2019. When to use parametric models in reinforcement learning? In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 14322–14333.

Wang, Z.; Schaul, T.; Hessel, M.; Van Hasselt, H.; Lanctot, M.; and De Freitas, N. 2016. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, 1995–2003.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021 [Spotlight Presentation]*.

Ye, W.; Liu, S.; Kurutach, T.; Abbeel, P.; and Gao, Y. 2021. Mastering Atari Games with Limited Data. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021,*.

Zanette, A.; Lazaric, A.; Kochenderfer, M. J.; and Brunskill, E. 2020. Learning Near Optimal Policies with Low Inherent Bellman Error. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 10978–10989. PMLR.