Hardness in Markov Decision Processes: Theory and Practice

Anonymous Author(s)
Affiliation
Address
email

Abstract

Meticulously analysing the empirical strengths and weaknesses of reinforcement learning methods in hard (challenging) environments is essential to inspire innovations and assess progress in the field. In tabular reinforcement learning, there is no well-established standard selection of environments to conduct such analysis, which is partially due to the lack of a widespread understanding of the rich theory of hardness of environments. The goal of this paper is to unlock the practical usefulness of this theory through four main contributions. First, we present a systematic survey of the theory of hardness, which also identifies promising research directions. Second, we introduce `Colosseum`, a pioneering Python package that enables empirical hardness analysis and implements a principled benchmark composed of environments that are diverse with respect to different measures of hardness. Third, we present an empirical comparison that provides new insights into current (efficiently computable) measures. Finally, we report the results of state-of-the-art tabular reinforcement learning algorithms in our newly proposed benchmark. Our contributions to tabular reinforcement learning are intended as solid steps towards the development of more principled benchmarks for the non-tabular setting.

1 Introduction

Reinforcement learning studies a setting where an agent interacts with an environment by observing states, receiving rewards, and selecting actions with the objective of optimizing a reward-based criterion. The field has attracted significant interest in recent years after striking performances obtained in board games [24] and video games [27, 6]. Solving these grand challenges constitutes an important milestone in the field. However, the corresponding agents require efficient simulators due to their high sample complexity (the number of observations that they require to optimize a reward-based criterion in an unknown environment). Outside of games, many important applications in healthcare, robotics, logistics, finance, and advertising can also be naturally formulated as reinforcement learning problems. However, simulators for these scenarios may not be available, reliable, or efficient.

The development of reinforcement learning methods that explore efficiently has long been considered (one of) the most crucial efforts to reduce sample complexity. Meticulously evaluating the strengths and weaknesses of such methods is essential to assess progress and inspire new developments in the field. Such empirical evaluations must be performed using *benchmarks* composed of a selection of environments and evaluation criteria. Ideally, this selection should be based on theoretically principled reasoning that considers the *hardness* of the environments and the soundness of the evaluation criteria.

In non-tabular reinforcement learning, where the number of states is large (and potentially infinite), there is no theory of hardness except for a few restricted settings. Consequently, the selection of environments in current benchmarks [21, 23] relies solely on the experience of their authors. Although such benchmarks are certainly valuable, there is no guarantee that they contain a sufficiently

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

diverse range of environments and that they are effectively able to quantify the agents' capabilities. In contrast, in tabular reinforcement learning, where the number of states and actions is finite, a rich theory of hardness of environments is available. Perhaps surprisingly, a principled benchmark based on this theory has so far been absent. There are at least two reasons for this absence. First, these hardness measures have been developed to provide theoretical guarantees for reinforcement learning methods, and not considered as directly useful for practical purposes. Second, the lack of a unifying presentation and critical comparison between different measures has limited their potential impact.

The goal of this paper is to establish the importance of hardness measures outside the context of theoretical reinforcement learning and unlock their practical usefulness. This goal is achieved through four main contributions. In Section 2, we present a systematic survey of the theory of hardness for Markov decision processes (MDPs). This survey serves as an introduction, tutorial, and, equally importantly, identifies gaps in the current theoretical landscape that suggest promising directions for future work. In Section 3.1, we introduce `Colosseum`, a pioneering Python package that enables the empirical investigation of MDP hardness and implements a principled benchmark for the four most widely studied tabular reinforcement learning settings. The environments in the benchmark were selected to maximize diversity with respect to two important measures of hardness, providing a varied set of challenges for which a *precise* characterization of hardness is available. In Section 3.2, we present an empirical comparison between three theoretical measures of hardness that can be computed efficiently. Our analysis provides insights into which aspects of hardness are best captured by each of the measures and identifies desirable qualities for future measures. Finally, in Section 3.3, we report the results of five agents with theoretical guarantees in our novel (principled) benchmark, empirically validating the hardness diversity of the selected environments. These agents are integrated into `Colosseum` and serve both as baselines and templates for implementing additional algorithms.

Although this paper is concerned with *tabular* reinforcement learning, for which principled measures of hardness are available, we believe our contributions are important milestones towards the future development of theoretical and empirical measures of hardness for non-tabular reinforcement learning.

2 Hardness in Theory

Section 2.1 introduces our notation and important definitions. Section 2.2 presents our survey of the theoretical landscape of measures of hardness, which includes a novel categorization of existing measures. Section 2.3 highlights the weaknesses of existing measures and introduces the concept of a *complete* measure of hardness, which we believe should be the focus of future developments.

2.1 Preliminaries

Let $\Delta(\mathcal{X})$ denote the set of probability distributions over a set $\mathcal{X}$. A finite Markov decision process M is a tuple $(\mathcal{S}, \mathcal{A}, P, P_0, R)$, where $\mathcal{S}$ is the finite set of states, $\mathcal{A}$ is the finite set of actions, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $P_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $R : \mathcal{S} \times \mathcal{A} \to \Delta([0, 1])$ is the reward kernel [22]. Given an optimization horizon T, a reinforcement learning agent aims to find a (possibly stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ that optimizes a reward-based criterion. The MDP M is unknown to the agent, but can be learned through experience. The interaction between agent and environment starts when the initial state $s_0 \sim P_0$ is drawn. For any $0 \leq t < T$, the agent samples an action $a_t \sim \pi_t(s_t)$ from its current policy π_t, and the environment draws the next state $s_{t+1} \sim P(s_t, a_t)$ and reward $r_{t+1} \sim R(s_t, a_t)$. An MDP is *episodic* with time horizon H when the state s_{t+1} is drawn from the initial state distribution whenever $t + 1 \equiv 0 \pmod{H}$ and *continuous* otherwise. In the continuous setting, a factor $\gamma \in (0, 1)$ is used to discount future rewards in the *discounted* setting, and future rewards are averaged across time in the *undiscounted* setting. For a policy π, an MDP M induces a Markov chain [17] with transition probabilities $\mathrm{P}^{\pi}_{s \to s'}(\mathrm{M}) = \sum_{a \in \mathcal{A}} \pi(a \mid s) P(s' \mid s, a)$ whose stationary distribution is denoted by μ^{π}. MDPs can be classified into three communication classes [13]. An MDP is *ergodic* if the Markov chain induced by any deterministic policy is ergodic. An MDP is *communicating* if, for every two states $s, s' \in \mathcal{S}$, there is a deterministic policy that induces a Markov chain where s is accessible from s' and vice versa. An MDP is *weakly communicating* when there is a partition $(\mathcal{C}, \mathcal{S} \setminus \mathcal{C})$ of the set of states $\mathcal{S}$ such that, for every two states $s, s' \in \mathcal{C}$, there is a deterministic policy that induces a Markov chain where s is accessible from s' and vice versa, and every state $s \in \mathcal{S} \setminus \mathcal{C}$ is transient in

every Markov chain induced by any deterministic policy. Every ergodic MDP is communicating, and every communicating MDP is weakly communicating.

The episodic state-action value function is given by $Q^{\pi}_{h,\texttt{epi}}(s,a) := \mathbb{E}\left[\sum_{t=h+1}^{H} r_t | s_h = s, a_h = a\right]$, and the episodic state value function is given by $V^{\pi}_{h,\texttt{epi}}(s) := \sum_a \pi(a \mid s) Q^{\pi}_{h,\texttt{epi}}(s,a)$. The discounted state-action value function is given by $Q^{\pi}_{\gamma}(s,a) := \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \mid s_0 = s, a_0 = a\right]$, and the discounted state value function is given by $V^{\pi}_{\gamma}(s) := \sum_a \pi(a \mid s) Q^{\pi}_{\gamma}(s,a)$. These expectations are taken with respect to the policy π, the transition kernel P, and the reward kernel R. In the undiscounted setting, the value of every state(-action) is the same, since rewards are averaged across infinite time steps. In that case, we define the expected average reward as $\rho^{\pi} := \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[r_t\right]$, which is also given by $\rho^{\pi} = \langle \mu^{\pi}, \mathrm{R}^{\pi} \rangle$, where R^{π} is a vector such that $\mathrm{R}^{\pi}_{s} = \mathbb{E}_{a\sim\pi(s)} R(s,a)$ is the expected reward obtained when following policy π from state s. An optimal policy π^* obtains maximum value for every state. We assume with little loss of generality that there is a unique optimal policy. We drop the corresponding subscripts when the (episodic or discounted) setting is clear from the context, and write $*$ to denote π^* in superscripts.

The most widely studied performance criteria are the *expected cumulative regret*, which yields the *regret minimization* setting [3], and the *sample efficiency of exploration*, which yields the Probably Approximately Correct Reinforcement Learning (PAC-RL) setting [12]. In simple terms, the regret measures the loss in reward due to the execution of a sub-optimal policy. In contrast, the sample efficiency measures how many interactions the agent requires to approximately learn π^* with high probability. The main difference between the two criteria is that the rewards obtained during the interaction with the MDP are not considered in PAC-RL, whereas all rewards contribute to the cumulative regret. Therefore, PAC-RL agents can afford more aggressive exploration.

2.2 Characterization of hardness

We identified two approaches for characterizing the hardness of an MDP in previous work. The first approach, which we call *Markov chain-based*, considers that an MDP is an extension of a Markov chain where transition probabilities can be changed based on direct interventions by an agent [5]. This approach is well suited to capture the hardness that comes from the *state visitation complexity*, which relates to the difficulty of visiting all the states in an MDP. Since learning tabular agents are required to visit every state at least once to identify the optimal policy, the state visitation complexity captures an important aspect of the hardness of an MDP. The second approach, which we call *value-based*, focuses on capturing the *estimation complexity*, which relates to the difficulty of estimating the optimal value function. Concretely, value-based measures consider the discrepancy between the values of neighboring states or how hard it is to act optimally given a value function estimate. Note that the two kinds of measures capture complementary aspects of MDP hardness. Markov chain-based measures may neglect the hardness induced by the reward kernel, whereas value-based measures may neglect the difficulty of obtaining the samples required to accurately estimate value functions.

Markov chain properties and value functions depend on a fixed policy, which presents two natural choices to derive measure hardness. The first choice considers the optimal policy, which typically leads to a measure that considers a best-case scenario. The second choice considers a policy that maximizes a criterion that characterizes a worst-case scenario. For instance, a policy that maximizes such criterion may spend its time in a region of the state space that is irrelevant for learning π^*.

2.2.1 Markov chain-based measures of hardness

Mixing time. The mixing time of a Markov chain with stationary distribution μ is defined as

$$t_{\mu} := \inf\{n \mid \sup_{s\in\mathcal{S}} d_{\mathrm{TV}}\left(p^n_s, \mu\right) \leq 0.25\}, \tag{1}$$

where d_{TV} is the total variation distance between distributions and the vector p^n_s represents the distribution over states after n steps starting from state s. For ergodic and aperiodic Markov chains, $\lim_{n\to\infty} p^n_s = \mu$ for every state s, so the mixing time is the number of steps a Markov chain takes to produce samples that are *close* to being distributed according to the stationary distribution μ. For instance, a Markov chain with a non-negligible probability of transitioning from every state to every state is quickly mixing. In contrast, the mixing time can be very long in chains where the state space has several distinct regions each of which is well connected but where transitions between regions

have low probability [15]. Kearns and Singh [14] propose an extension of the mixing time to MDPs that considers the maximum mixing time across policies $t := \sup_\pi t_{\mu^\pi}$ in the undiscounted setting. Although the mixing time plays an important role in that setting, since the average reward obtained by policy π is given by $\rho^\pi = \langle \mu^\pi, \mathrm{R}^\pi \rangle$, it is not a generally good measure of hardness. First, it does not capture the state visitation complexity, since it neglects the fact that the agent may direct its exploration through a choice of policy. Second, it does not capture any significant aspect of value function estimation, which does not require samples from the stationary distribution of a given policy.

Diameter. The diameter is fundamentally related to the number of time steps required to transition between states. In the continuous setting, the diameter D is most commonly defined as

$$D := \sup_{s_1 \neq s_2} \inf_\pi T^\pi_{s_1 \to s_2}, \tag{2}$$

where $T^\pi_{s_1 \to s_2}$ is the expected number of time steps required to reach state s_2 from state s_1 when following policy π [10]. Intuitively, D is the worst-case expected number of time steps required to transition between two states when following the best policy for that purpose. Related definitions are $D_{\text{worst}} := \sup_\pi \sup_{s_1 \neq s_2} T^\pi_{s_1 \to s_2}$ and $D_{\text{opt}} := \inf_\pi \sup_{s_1 \neq s_2} T^\pi_{s_1 \to s_2}$, which imply $D \leq D_{\text{opt}} \leq D_{\text{worst}}$ [4].

In the episodic setting, we define the diameter by augmenting each state s with the current (in-episode) time step h and considering the diameter of this augmented MDP in the continuous setting. Note that $T^\pi_{(s_1,h_1) \to (s_2,h_2)}$ can be larger than the episode length H, which means that (on average) more than one episode may be required to transition from state (s_1, h_1) to state (s_2, h_2). This happens whenever $h_2 < h_1$, which may be undesirable if the intent is to focus on the expected number of time steps required to transition between states from the same episode. The diameter is always infinite in weakly-communicating MDPs if the supremum is not restricted to states in the recurrent class $\mathcal{C}$ and is always finite in the episodic setting (where every state is reachable within an episode). A large diameter can be caused by high stochasticity. The diameter is very apt at measuring state visitation complexity, since it captures the effort required to deliberately move between states. However, it neglects the reward kernel, and so has limited capacity to measure estimation complexity.

Distribution mismatch coefficient. The *distribution mismatch coefficient* (DMC) has been defined for the continuous undiscounted [29] and continuous discounted [1] cases respectively as

$$\text{DMC} := \sup_\pi \sum_{s \in \mathcal{S}} \frac{\mu^*_s}{\mu^\pi_s} \qquad \text{and} \qquad \text{DMC}_{s_0} := \sup_{s \in \mathcal{S}} \frac{d^*_{s_0}(s)}{P_0(s)},$$

where $d^*_{s_0}(s) = (1-\gamma) \sum_{t=0}^\infty \gamma^t \text{Pr}(s_t = s \mid s_0)$ is the discounted state visitation distribution of the optimal policy given an initial state s_0. Note that the DMC is guaranteed to be finite only for ergodic MDPs. In communicating MDPs, there is at least one policy π whose stationary distribution assigns probability zero to some states. MDPs whose optimal stationary distribution μ^* has its probability mass concentrated on a few states tend to have a large DMC. In contrast, when μ^* is closer to being uniformly distributed across states, the DMC tends to be small. For small values of DMC, as every μ^π is *close* to μ^*, the agent will gather samples from the optimal stationary distribution regardless of its current policy, which may enable quick learning. In contrast, for large values of DMC, the agent needs to actively seek a policy that gathers such samples. The DMC is not well suited to quantify the state visitation complexity, since it fails to capture the difficulty of visiting all states. The DMC also does not capture the estimation complexity, since it does not account for the stochasticity of the environment, which is related to the number of samples required to make correct estimations.

2.2.2 Value-based measures of hardness

Action-gap regularity. Given an estimate $\hat{Q}^*$ of the optimal state-action value function Q^*, the *greedy policy* with respect to $\hat{Q}^*$ always chooses an action associated with the highest state-action value. Whether or not such a policy is optimal depends exclusively on the ordering of the estimates for a given state rather than their accuracy. For instance, assuming that a^* is the optimal action for every state s, a greedy agent would act optimally if $\hat{Q}(s, a^*) > \hat{Q}(s, a')$ for every action $a' \neq a^*$, even if $|Q^*(s,a) - \hat{Q}(s,a)| \gg 0$ for every action a. The action-gap regularity ζ is a measure of hardness that leverages this principle through the theory of hardness for classification algorithms [9]. However, this measure is only defined for two actions, and so has exceptionally limited applicability.

Environmental value norm. The (discounted) environmental value norm C_γ^π is defined as

$$C_\gamma^\pi := \sup_{(s,a)} \sqrt{\operatorname*{Var}_{s' \sim P(s,a)} V^\pi(s')}.$$

This quantity can be similarly defined in the undiscounted setting [18]. In words, the environmental value norm captures the one-step variance of the value function V^π for a given policy π. In the episodic setting, the analogous *maximum per-step conditional variance* C_{H}^π [30] is defined as

$$C_{\mathrm{H}}^\pi := \sup_{(s,a,h)} \left(\operatorname{Var} R(s,a) + \operatorname*{Var}_{s' \sim P(s,a)} V_{h+1}^\pi(s') \right).$$

Alternatively, as with the diameter, it is also possible to define this quantity by augmenting each state s with the current (in-episode) time step h and considering the environmental value norm of this augmented MDP in the continuous setting. For every policy, the environmental value norm is equal to zero when the transition kernel is deterministic. However, a highly stochastic MDP may still have a small environmental value norm, since this norm captures the variance of the state value function rather than the stochasticity of the transition kernel. Maillard et al. [18] suggest using the environmental value norm of the optimal policy π^* as a measure of hardness. The main strength of such measure is that the variability of the optimal value function captures an important aspect of the estimation complexity. However, this measure of hardness neglects the state visitation complexity.

Sub-optimality gap. The sub-optimality gap is defined in the continuous and episodic settings as

$$\Delta(s,a) := V^*(s) - Q^*(s,a) \quad \text{and} \quad \Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a),$$

respectively. Since $V^*(s) = \max_a Q^*(s,a)$ for every state s, the sub-optimality gap $\Delta(s,a)$ measures the difference in expected return between selecting the optimal action for state s and selecting the action a. Intuitively, identifying a suboptimal action a' in a given state s is easier if the gap $\Delta(s,a')$ is large. Simchowitz and Jamieson [25] identifies the sum of the reciprocals of the sub-optimality gaps $\sum_{(s,a)|\Delta(s,a)\neq 0} \frac{1}{\Delta(s,a)}$ as a measure of hardness. They also demonstrate the importance of the sub-optimality gaps by showing that recent optimistic algorithms necessarily incur in a cumulative regret proportional to the smallest nonzero sub-optimality gap in the episodic setting. However, note that approximating the optimal value function (and identifying a near-optimal policy) is particularly easy when every sub-optimality gap is small. Consequently, the (PAC-RL) sample complexity is likely to decrease when the sum of the reciprocals of the sub-optimality gaps increases. Furthermore, this measure does not explicitly capture visitation complexity and is prone to severe numerical issues.

2.3 Complete measures of hardness

As shown through the previous section, current measures of hardness are limited in their ability to simultaneously capture state visitation complexity and estimation complexity. Additionally, current measures are also oblivious to the distinct challenges presented by the two most widely studied performance criteria. For instance, consider an MDP where a set of lowly rewarding states stands between an agent and a set of highly rewarding states. A regret minimization agent is heavily discouraged from visiting the lowly rewarding states, in contrast to a PAC-RL agent. Therefore, the state visitation complexity should differ across settings. The fact that current measures disregard this distinction is concerning, since they should quantify the difficulty of a specific optimization task.

Both issues are not discussed in previous work, since measures of hardness have not been considered relevant outside the context of deriving theoretical performance guarantees for reinforcement learning agents. In order to address these issues in the theory of hardness, we propose the following definition.

Definition 2.1 (Complete measure of hardness) A measure of hardness $\theta : \mathcal{M} \to \mathbb{R}^+$ is complete for an MDP class $\mathcal{M}$ and criterion ψ (sample complexity or expected cumulative regret) if, for every pair of MDPs $\mathrm{M}_1, \mathrm{M}_2 \in \mathcal{M}$ and *near-optimal agent* A^* that achieves the criterion lower bound of class $\mathcal{M}$ up to logarithmic factors, $\theta(\mathrm{M}_1) > \theta(\mathrm{M}_2)$ implies $\psi(\mathrm{M}_1, A^*) > \psi(\mathrm{M}_2, A^*)$.

The development of efficiently computable (non-trivial) complete measures of hardness is a promising direction for future work, and combining existing measures that capture state visitation complexity and estimation complexity is a viable first step in that direction. Recently, Wagenmaker et al. [28]

have pioneered this approach in the episodic setting by proposing the *gap-visitation complexity*

$$\text{GVP}(\epsilon) := \sum_{h=0}^{H} \inf_{\pi} \sup_{(s,a)} \inf \left(\frac{1}{w_h^\pi(s,a)\Delta_h(s,a)^2}, \frac{W_h(s)^2}{w_h^\pi(s,a)\epsilon^2} \right), \tag{3}$$

where $w_h^\pi(s,a) := P_\pi(s_h = s, a_h = a)$ is the probability of visiting state-action pairs (s,a) at (in-episode) time step h when following policy π, $W_h(s) := \sup_\pi P_\pi(s_h = s)$ is the maximum reachability of state s at (in-episode) time step h, and ϵ is a parameter related to the optimality of the output policy in the PAC-RL setting. The strength of this measure is that it weights sub-optimality action gaps with measures of visitation complexity (w_h^π and W_h). This is important because the presence of a state action pair that is both hard to estimate and hard to reach significantly impacts the hardness of an MDP. However, as a measure conceived to provide theoretical guarantees in the PAC-RL setting, the gap-visitation complexity is dependent on a choice of ϵ, restricted to the finite horizon setting, not efficiently computable, and prone to numerical issues.

3 Hardness in Practice

Section 3.1 introduces the `Colosseum` package. Section 3.2 investigates measures of hardness empirically. Section 3.3 reports the results of near-optimal agents in our newly proposed benchmark.

3.1 Colosseum

`Colosseum` bridges theory and practice in tabular reinforcement learning by being the first tool to provide a principled benchmark and allow empirically investigating the hardness of MDPs.[1]

As an analysis tool, `Colosseum` provides several functionalities: it implements measures of hardness such as the diameter (App. C describes our novel solution to compute this measure), the sum of the reciprocals of the sub-optimality gaps, and the environmental value norm; it computes the expected cumulative regret in the episodic and continuous settings; it provides the first implementation of an algorithm to identify the communication class of an MDP [13]; it assembles insightful visualizations and logs of interactions between agents and MDPs; and it provides a simple interface that allows implementing MDPs beyond the eight families already included (App. B details these families).

As a benchmarking tool, `Colosseum` is unique in its strong connection with theory. For instance, in contrast to non-tabular benchmarks, the `Colosseum` benchmark allows computing theoretical evaluation criteria such as the expected cumulative regret. The benchmark covers the most commonly studied MDP classes: *episodic ergodic*, *episodic communicating*, *continuous ergodic*, and *continuous communicating*. From each class, we have selected twenty MDPs that are as diverse as possible with respect to their diameters and environmental value norms (Fig. 16 in App. F shows how each of these MDPs varies according to these measures and Sec. 3.3 further validates this selection), which represent different combinations of state visitation complexity and estimation complexity. Notably, this makes the `Colosseum` benchmark the most exhaustive in tabular reinforcement learning, since previous empirical evaluations are conducted in few MDPs (such as Taxi [8] or RiverSwim [26]).

Appendix A provides more details about `Colosseum` design choices, structure, and procedures.

3.2 Empirical analysis of hardness measures

For brevity, this section only presents results of hardness measures in the `MiniGridEmpty` family of environments in the episodic setting. Appendix E presents results for additional families.

A `MiniGridEmpty` MDP is a grid world where an agent has three available actions: moving forward, rotating left, and rotating right. An agent is rewarded for being in a few specific states and receives no reward in every other state. Appendix B.4 provides more details about this family of environments.

In our investigation, we consider four scenarios that highlight the different aspects of hardness.

Scenario 1. We vary the probability `p_rand` that an MDP executes a random action instead of the action selected by an agent. As `p_rand` approaches one, value estimation becomes easier, since outcomes depend less on choices. However, intentionally visiting states becomes harder.

[1] The (anonymized) open-source implementation is available on `https://anonymous.4open.science/r/Colosseum-B725/README.md`. A Jupyter Notebook tutorial is also provided as supplementary material.

Scenario 2. We vary the probability `p_lazy` that an MDP stays in the same state instead of executing the action selected by an agent. Contrary to increasing `p_rand`, increasing `p_lazy` never benefits exploration. Increasing `p_lazy` decreases estimation complexity and increases visitation complexity.

Scenario 3 and 4. We vary the number of states across MDPs from the same family. In scenario 4, we also let `p_rand` $= 0.1$ to study the impact of stochasticity. In these scenarios, increasing the number of states increases both estimation complexity and visitation complexity.

In every scenario, hardness measures are compared with the cumulative regret of a *near-optimal* agent tuned for each specific MDP (see App. E). This regret serves as an optimistic measure of hardness. Appendix D describes how these measures are normalized, which *solely* allows comparing trends.

Analysis. Figure 1 presents the empirical results for the episodic `MiniGridEmpty` family in the four scenarios with 95% bootstrapped confidence intervals over twelve random seeds.

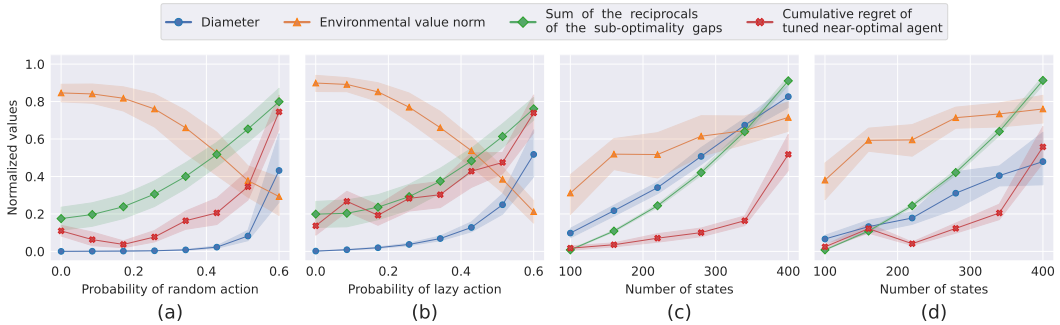

Figure 1: Episodic `MiniGridEmpty` results, scenarios 1-4 (subfigures a-d).

Diameter grows superlinearly with both `p_rand` and `p_lazy`, since deliberate movement between states requires an exponentially increasing number of time steps. Although the diameter highlights this sharply increasing visitation complexity, its trend overestimates the increase in cumulative regret of the tuned near-optimal agent, which is explained by the unaccounted decrease in estimation complexity. The diameter also increases almost linearly with the number of states. When `p_rand` is relatively small, an approximately linear relationship can still be observed. This linear trend underestimates the non-linear growth in hardness clearly shown in the cumulative regret of the tuned near-optimal agent, but is in line with the mild increase in visitation complexity. These findings corroborate our claim that the diameter captures visitation rather than estimation complexity.

Environmental value norm decreases as `p_lazy` and `p_rand` increase because the optimal value of neighboring states becomes closer, which decreases the per-step variability of the optimal value function. When the number of states increases but the transition and reward structures remain the same, the small increase in this variability only causes this measure to grow sublinearly. These findings are strong evidence that this measure is only suited to capture estimation complexity.

Sum of the reciprocals of the sub-optimality gaps increases weakly superlinearly in scenarios 1 and 2. The probability of executing the action selected by the agent decreases when `p_lazy` and `p_rand` increase, and so the difference between the optimal value function and the optimal state-action value function decreases sharply. The sum of the reciprocals of the sub-optimality gaps increases almost linearly with the number of states. This is explained by the fact that the average value of the additional terms in the summation is often similar to the average value of the existing terms given the same structure of reward and transition kernels. This measure of hardness is not particularly apt at capturing estimation complexity, since it focuses solely on optimal policy identification. It also underestimates the increase in hardness induced by an increase in visitation complexity.

3.3 The Colosseum benchmark

This section presents the results on the `Colosseum` benchmark of five diverse agents with known theoretical guarantees: posterior sampling for reinforcement learning (PSRL) for the episodic and continuous settings [19, 2], Q-learning with UCB exploration for the episodic setting [11], Q-learning with optimism for the continuous setting [29], and UCRL2 for the continuous setting [10]. Besides being valuable on their own, these experiments help validate our selection of environments.

We set the total number of time steps to $500\,000$ with a maximum training time of 20 minutes for the episodic case and 60 minutes for the continuous case. If an agent does not reach the maximum number of time steps before this time limit, learning is interrupted but the agent keeps using its last available policy. This guarantees that computationally expensive (model-based) methods can compete fairly with more computationally inexpensive (model-free) methods. More time is allocated to the continuous case because the corresponding algorithms are naturally more expensive. Each full interaction between an agent and an MDP is repeated with five seeds. The hyperparameters for the agents have been obtained as the result of a random search aiming to minimize the average cumulative regret across a set of MDPs of varying sizes and parameters (detailed in App. F). These MDPs are not part of the benchmark. The cumulative regret and the the per-step normalized cumulative regret (defined in App. D) are employed as performance measures. In Table 1, we report the per-step normalized regrets divided by the total number of time steps with standard deviations. For brevity, we describe the main findings in this section and leave an in-depth analysis for Appendix F.

Table 1: Normalized cumulative regrets of selected agents on the `Colosseum` benchmark. (a) Episodic ergodic. (b) Episodic communicating. (c) Continuous ergodic. (d) Continuous communicating.

(a)

MDP	Q-learning	PSRL
DeepSea	$.61 \pm .00$	$\mathbf{.02} \pm .00$
	$.51 \pm .00$	$\mathbf{.02} \pm .00$
FrozenLake	$.90 \pm .01$	$\mathbf{.01} \pm .00$
MG-Empty	$1.00 \pm .00$	$\mathbf{.21} \pm .04$
	$1.00 \pm .00$	$\mathbf{.32} \pm .09$
	$1.00 \pm .00$	$\mathbf{.30} \pm .05$
	$.99 \pm .00$	$\mathbf{.13} \pm .03$
	$.99 \pm .00$	$\mathbf{.19} \pm .05$
	$.92 \pm .05$	$\mathbf{.03} \pm .01$
	$.93 \pm .02$	$\mathbf{.01} \pm .00$
MG-Rooms	$.84 \pm .04$	$\mathbf{.03} \pm .00$
	$1.00 \pm .00$	$\mathbf{.33} \pm .04$
	$.99 \pm .00$	$\mathbf{.17} \pm .02$
RiverSwim	$.04 \pm .01$	$\mathbf{.00} \pm .00$
	$.90 \pm .01$	$\mathbf{.01} \pm .00$
SimpleGrid	$.77 \pm .03$	$\mathbf{.02} \pm .00$
	$.79 \pm .03$	$\mathbf{.18} \pm .01$
	$.49 \pm .03$	$\mathbf{.07} \pm .01$
Taxi	$.84 \pm .01$	$\mathbf{.05} \pm .00$
	$.56 \pm .02$	$\mathbf{.03} \pm .00$
Average	$.80 \pm .01$	$\mathbf{.11} \pm .02$

(b)

MDP	Q-learning	PSRL
DeepSea	$.01 \pm .00$	$\mathbf{.00} \pm .00$
	$.83 \pm .00$	$\mathbf{.28} \pm .04$
FrozenLake	$.79 \pm .03$	$\mathbf{.00} \pm .00$
MG-Empty	$.53 \pm .04$	$\mathbf{.07} \pm .04$
	$.99 \pm .00$	$\mathbf{.10} \pm .05$
	$1.00 \pm .00$	$\mathbf{.15} \pm .04$
	$1.00 \pm .00$	$\mathbf{.23} \pm .04$
	$1.00 \pm .00$	$\mathbf{.36} \pm .07$
MG-Rooms	$.99 \pm .00$	$\mathbf{.12} \pm .02$
	$1.00 \pm .00$	$\mathbf{.25} \pm .03$
	$1.00 \pm .00$	$\mathbf{.30} \pm .02$
	$.93 \pm .04$	$\mathbf{.04} \pm .01$
RiverSwim	$.85 \pm .00$	$\mathbf{.01} \pm .00$
	$.97 \pm .00$	$\mathbf{.02} \pm .00$
SimpleGrid	$.75 \pm .10$	$\mathbf{.07} \pm .01$
	$.79 \pm .00$	$\mathbf{.13} \pm .00$
	$.50 \pm .00$	$\mathbf{.02} \pm .00$
	$.78 \pm .04$	$\mathbf{.62} \pm .02$
Taxi	$.94 \pm .01$	$\mathbf{.05} \pm .01$
	$.90 \pm .01$	$\mathbf{.14} \pm .01$
Average	$.83 \pm .01$	$\mathbf{.15} \pm .02$

(c)

MDP	Q-learning	PSRL	UCRL2
DeepSea	$.92 \pm .00$	$.93 \pm .00$	$\mathbf{.38} \pm .01$
FrozenLake	$.82 \pm .03$	$.78 \pm .14$	$\mathbf{.08} \pm .01$
MG-Empty	$.95 \pm .00$	$.97 \pm .00$	$\mathbf{.39} \pm .02$
	$.94 \pm .01$	$.97 \pm .02$	$\mathbf{.29} \pm .02$
	$.94 \pm .00$	$.96 \pm .01$	$\mathbf{.57} \pm .01$
	$.96 \pm .00$	$.97 \pm .00$	$\mathbf{.67} \pm .02$
	$.87 \pm .01$	$.91 \pm .12$	$\mathbf{.11} \pm .00$
	$.96 \pm .00$	$.97 \pm .00$	$\mathbf{.49} \pm .24$
	$.94 \pm .01$	$.97 \pm .01$	$\mathbf{.31} \pm .04$
	$.97 \pm .00$	$.96 \pm .01$	$\mathbf{.73} \pm .04$
MG-Rooms	$.82 \pm .09$	$.98 \pm .00$	$\mathbf{.81} \pm .26$
	$.96 \pm .00$	$.98 \pm .00$	$\mathbf{.88} \pm .19$
RiverSwim	$.74 \pm .00$	$.27 \pm .29$	$\mathbf{.01} \pm .00$
	$.64 \pm .01$	$.57 \pm .36$	$\mathbf{.01} \pm .00$
	$.79 \pm .00$	$.36 \pm .41$	$\mathbf{.02} \pm .00$
	$.02 \pm .00$	$.82 \pm .21$	$\mathbf{.01} \pm .00$
SimpleGrid	$.76 \pm .00$	$.77 \pm .01$	$\mathbf{.24} \pm .01$
	$\mathbf{.07} \pm .00$	$.09 \pm .10$	$.10 \pm .13$
	$.46 \pm .00$	$.49 \pm .01$	$\mathbf{.29} \pm .16$
Taxi	$\mathbf{.76} \pm .01$	$.86 \pm .05$	$.78 \pm .01$
Average	$.76 \pm .01$	$.78 \pm .09$	$\mathbf{.36} \pm .06$

(d)

MDP	Q-learning	PSRL	UCRL2
DeepSea	$\mathbf{.74} \pm .00$	$.76 \pm .02$	$.86 \pm .02$
	$.97 \pm .00$	$\mathbf{.97} \pm .00$	$.97 \pm .00$
	$\mathbf{.75} \pm .00$	$.76 \pm .05$	$.86 \pm .03$
FrozenLake	$.75 \pm .03$	$.64 \pm .30$	$\mathbf{.04} \pm .01$
	$.86 \pm .05$	$.64 \pm .29$	$\mathbf{.18} \pm .16$
MG-Empty	$.69 \pm .02$	$.98 \pm .00$	$\mathbf{.04} \pm .01$
	$\mathbf{.01} \pm .00$	$.98 \pm .00$	$.01 \pm .00$
	$\mathbf{.00} \pm .00$	$.98 \pm .00$	$.00 \pm .00$
	$.77 \pm .01$	$.98 \pm .00$	$\mathbf{.03} \pm .00$
	$.87 \pm .01$	$.98 \pm .00$	$\mathbf{.43} \pm .43$
MG-Rooms	$\mathbf{.01} \pm .00$	$.98 \pm .00$	$.84 \pm .03$
	$\mathbf{.01} \pm .00$	$.98 \pm .00$	$.04 \pm .01$
	$\mathbf{.01} \pm .01$	$.98 \pm .00$	$.87 \pm .06$
RiverSwim	$.40 \pm .01$	$.16 \pm .31$	$\mathbf{.00} \pm .00$
	$.69 \pm .01$	$.09 \pm .03$	$\mathbf{.01} \pm .01$
SimpleGrid	$\mathbf{.09} \pm .01$	$.91 \pm .00$	$.46 \pm .35$
	$.00 \pm .00$	$.49 \pm .00$	$\mathbf{.00} \pm .00$
	$\mathbf{.18} \pm .03$	$.91 \pm .00$	$.90 \pm .02$
	$.01 \pm .01$	$.49 \pm .00$	$\mathbf{.01} \pm .00$
Taxi	$\mathbf{.87} \pm .01$	$.93 \pm .04$	$.92 \pm .03$
Average	$.43 \pm .01$	$.78 \pm .05$	$\mathbf{.37} \pm .06$

Analysis. Table 1 often shows significant variability in the performance of the same agent across MDPs of the same family. Therefore, the criteria employed to select MDPs for the benchmark have yielded diverse challenges even in the case of similar transition and reward structures. For example, in the continuous communicating case (Tab. 1d), Q-learning finds the optimal policy for some MDPs in the `MiniGridEmpty` family but fails for others. This also happens with UCRL2 in the `SimpleGrid`.

The ergodic settings seem generally slightly easier than the communicating settings. Exceptionally, in the continuous setting, the ergodic setting is more challenging than the communicating setting for Q-learning (Tabs. 1c and 1d). In `Colosseum`, the ergodic MDPs are obtained by letting $\texttt{p_rand} > 0$ in otherwise communicating MDPs. Designing a naturally ergodic MDP is not straightforward, and so the vast majority of MDPs employed in the literature are communicating. However, model-free algorithms struggle with the corresponding increased variability of the state-action value function.

In the episodic settings (Tabs. 1a and 1b), PSRL obtains excellent results and consistently solves all the MDPs with a remarkably small variability. In contrast, Q-learning only finds the optimal policy for a few MDPs. This often happens because Q-values are drastically updated when an MDP executes an action with a large sub-optimality gap instead of a good action chosen by an agent, which requires a significant number of samples to be overcome for optimal action selection.

In the continuous settings (Tabs. 1c and 1d), UCRL2 is clearly the best algorithm. However, it sometimes presents a relatively high standard deviation of cumulative regret. In contrast to the episodic setting, PSRL struggles to find the optimal policy for the majority of the MDPs and generally obtains the worst results. The reason for its weak performance in this setting is the computationally expensive optimistic sampling procedure that is required to obtain theoretical guarantees. In practice, PSRL almost always breaks the time limit before reaching the first quarter of the total number of time steps, before gathering enough data to find an optimal policy. Furthermore, it is too optimistic to efficiently solve non-pathological MDPs.

In order to illustrate how hardness measures relate with cumulative regret, Figure 2 places the average cumulative regret obtained by each agent in each continuous ergodic MDP in a coordinate that corresponds to the diameter and the environmental value norm of that MDP. For UCRL2 (Fig. 2a), there is generally a positive relationship between both of these hardness measures and the average cumulative regret. For Q-learning and PSRL (Figs. 2b and 2c), the diameter seems to have a generally smaller influence on the average cumulative regret. Appendix F presents analogous figures for the remaining settings, which are intended to further validate our selection of MDPs for the benchmark.

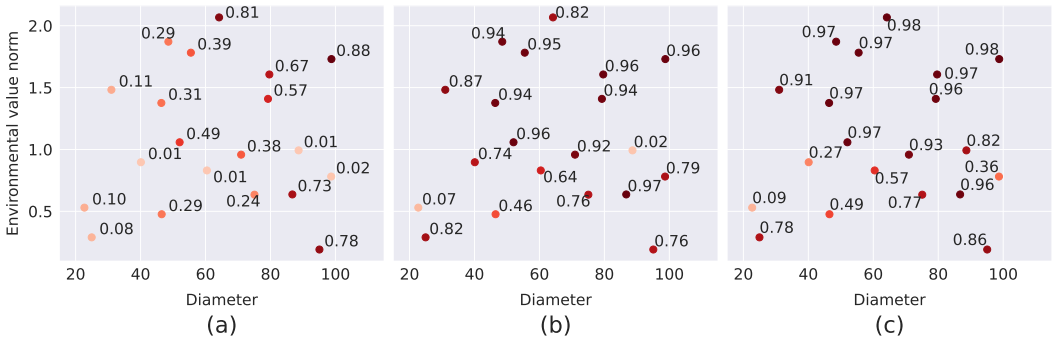

Figure 2: Average cumulative regret obtained by each agent in each continuous ergodic MDP, placed according to two hardness measures associated to that MDP. (a) UCRL2. (b) Q-learning. (c) PSRL.

4 Conclusion

We established the usefulness of hardness measures in empirical reinforcement learning. Prior to our work, the application of these measures was limited to the development of theoretical performance guarantees for reinforcement learning agents. In order to encourage a widespread understanding of these measures, we presented a systematic survey that newly identified two major approaches for characterizing hardness: Markov-chain based and value-based. These approaches capture two main aspects of hardness: visitation complexity and estimation complexity. Our survey also exposed a relative lack of measures that capture both aspects, which motivates our definition of complete measures of hardness. The development of such measures is theoretically and empirically important.

We have also presented the first empirical study of (efficiently computable) hardness measures using several families of environments found in previous work. This study revealed how existing measures capture different aspects of hardness and clarified their relationship with the behavior of near-optimal agents. Based on these results, we proposed a benchmark for the most widely studied tabular reinforcement learning settings that contains environments that maximize diversity with respect to two highly distinct measures. Such a principled benchmark is invaluable to gauge progress in the field. The new benchmark allowed conducting the most exhaustive empirical comparison between theoretically principled tabular reinforcement learning agents to date, which revealed previously undocumented weaknesses of these agents and further validated our choices of environments.

Our work has lead to the development of `Colosseum`, a pioneering tool for the empirical study of tabular reinforcement learning. Besides implementing the aforementioned benchmark, `Colosseum` provides invaluable analysis tools: regret and hardness computations, communication class identification, logging, and visualizations. `Colosseum` can also be easily extended and integrated with new environments. This open-source tool will be publicly available, and we will actively seek contributions to implement more agents and environments. We strongly believe that `Colosseum` has the potential to become a fundamental tool in reinforcement learning.

We envision many possibilities for future work. They include the development of complete measures of hardness; the creation of measures that capture both state visitation complexity and estimation complexity; the identification of new aspects of hardness; and, not less importantly, the development of theoretical and empirical measures of hardness for non-tabular reinforcement learning.

References

[1] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 2021.

[2] S. Agrawal and R. Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *Advances in Neural Information Processing Systems*, 2017.

[3] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 2006.

[4] P. Bartlett and A. Tewari. REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence: Proceedings of the 25th Conference*. AUAI Press, 2009.

[5] R. Bellman. A Markovian decision process. *Journal of mathematics and mechanics*, 1957.

[6] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

[7] M. Chevalier-Boisvert, L. Willems, and S. Pal. Minimalistic Gridworld environment for OpenAI gym. `https://github.com/maximecb/gym-minigrid`, 2018.

[8] T. G. Dietterich. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of artificial intelligence research*, 2000.

[9] A. M. Farahmand. Action-gap phenomenon in reinforcement learning. *Advances in Neural Information Processing Systems*, 2011.

[10] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

[11] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? *Advances in neural information processing systems*, 2018.

[12] S. M. Kakade. *On the sample complexity of reinforcement learning*. University of London, University College London (United Kingdom), 2003.

[13] L. Kallenberg. Classification problems in MDPs. In *Markov processes and controlled Markov chains*. Springer, 2002.

[14] M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 2002.

[15] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. Adaptive Computation and Machine Learning series. MIT Press, 2009.

[16] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A LLVM-based python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, 2015.

[17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. American Mathematical Soc., 2017.

[18] O.-A. Maillard, T. A. Mann, and S. Mannor. How hard is my MDP? The distribution-norm to the rescue". *Advances in Neural Information Processing Systems*, 2014.

[19] I. Osband, D. Russo, and B. Van Roy. (More) Efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 2013.

[20] I. Osband, B. Van Roy, D. J. Russo, Z. Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 2019.

[21] I. Osband, Y. Doron, M. Hessel, J. Aslanides, E. Sezener, A. Saraiva, K. McKinney, T. Lattimore, C. Szepesvári, S. Singh, B. Van Roy, R. Sutton, D. Silver, and H. van Hasselt. Behaviour suite for reinforcement learning. In *International Conference on Learning Representations*, 2020.

[22] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[23] R. Rajan, J. L. B. Diaz, S. Guttikonda, F. Ferreira, A. Biedenkapp, J. O. von Hartz, and F. Hutter. Mdp playground: A design and debug testbed for reinforcement learning. *arXiv preprint arXiv:1909.07750*, 2019.

[24] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 2018.

[25] M. Simchowitz and K. G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. *Advances in Neural Information Processing Systems*, 2019.

[26] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 2008.

[27] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 2019.

[28] A. Wagenmaker, M. Simchowitz, and K. Jamieson. Beyond no regret: instance-dependent PAC reinforcement learning. *arXiv preprint arXiv:2108.02717*, 2021.

[29] C.-Y. Wei, M. J. Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International conference on machine learning*. PMLR, 2020.

[30] A. Zanette and E. Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In the introduction, we explain where each contribution is described in the text. The same contributions are mentioned in the abstract. In the conclusion, we also present a concrete list of contributions.
   (b) Did you describe the limitations of your work? [Yes] We describe the assumptions behind every theoretical measure of hardness. The results of our empirical investigation are naturally limited to the environments, measures, and agents that we considered. Nevertheless, we believe that the experimental protocol described in Appendices E and F is very robust. The remaining limitations are discussed as potential future works.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We do not envision potential negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We explicitly mention the full set of assumptions involved in theoretical results and report relevant references when appropriate.
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include a link to an anonymized version of the code in Section 3. The same code is also included as supplementary material together with a Jupyter Notebook tutorial. We describe the

experimental procedures for the empirical evaluation of the measures of hardness in Section 3.2 and for the `Colosseum` benchmark evaluation in Section 3.3. We refer to the corresponding appendices (Apps. E and F) for additional details.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] The details for the empirical evaluation of the measures of hardness and the `Colosseum` benchmark evaluation are provided in Appendices E and F.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We report 95% bootstraped confidence intervals in the plots and standard deviations in the tables.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] These details are provided in Appendices E and F.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We describe all the dependencies of `Colosseum` in Appendix A.

(b) Did you mention the license of the assets? [No] However, all these assets are open-source and freely available.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code of our newly proposed Python package (`Colosseum`) is included in a link in Section 3 and in the supplementary material.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]