Continual Optimistic Initialization for Value-Based Reinforcement Learning

Sheelabhadra Dey  
Texas A&M University  
College Station, USA  
sheelabhadra@tamu.edu

James Ault  
Texas A&M University  
College Station, USA  
jault@tamu.edu

Guni Sharon  
Texas A&M University  
College Station, USA  
guni@tamu.edu

ABSTRACT

Comprehensive state-action exploration is essential for reinforcement learning (RL) algorithms. It enables them to find optimal solutions and avoid premature convergence. In value-based RL, optimistic initialization of the value function ensures sufficient exploration for finding the optimal solution. Optimistic values lead to curiosity-driven exploration enabling visitation of under-explored regions. However, optimistic initialization has limitations in stochastic and non-stationary environments due to its inability to explore “infinitely-often”. To address this limitation, we propose a novel exploration strategy for value-based RL, denoted COIN, based on recurring optimistic initialization. By injecting a continual exploration bonus, we overcome the shortcoming of optimistic initialization (sensitivity to environment noise). We provide a rigorous theoretical comparison of COIN versus existing popular exploration strategies and prove it provides a unique set of attributes (coverage, infinite-often, no visitation tracking, and curiosity). We demonstrate the superiority of COIN over popular existing strategies on a designed toy domain as well as present results on common benchmark tasks. We observe that COIN outperforms existing exploration strategies in four out of six benchmark tasks while performing on par with the best baseline on the other two tasks.

KEYWORDS

Reinforcement Learning; Exploration Strategies; Optimistic Initialization

ACM Reference Format:

1 INTRODUCTION

Sequential decision problems are commonly modeled as Markov decision processes (MDP) [49]. A solution to an MDP is a policy that maps states to actions such that the resulting sequential behavior is optimal with respect to a given utility function. Reinforcement learning (RL) algorithms [61] are designed to solve MDPs via interactions with the underlying environment. RL algorithms were shown to be effective on a variety of applications such as robotics [2, 37], autonomous driving [25, 29, 36], traffic signal control [9, 10, 72], and drilling operations [5, 35, 68]. An RL agent commonly explores the environment by executing actions and observing outcomes. The acquired knowledge enables the RL agent to reason about the optimal policy. RL algorithms commonly require that the Cartesian product of the state and action spaces is sufficiently explored. Consequently, Exploration strategies are usually designed to provide some theoretical assurances regarding the state-action visitation distribution. These assurances result in desirable learning guarantees [7, 8, 26, 60] relating to, policy convergence, policy optimality, speed of convergence (sample efficiency), memory complexity, and regret bounds. Existing exploration strategies can be divided into three categories (following taxonomy by Amin et al. [4]) (C1) sampling-based, (C2) visitation-tracking, and (C3) optimistic initialization-based. Each of these classes has unique benefits and drawbacks with respect to the following properties: (P1) state-action space coverage, (P2) infinite-often visitation, (P3) visitation tracking, and (P4) curiosity.

In this paper, we first study the effectiveness of exploration strategies in a specially designed MDPs denoted bridge crossing problems (explained later in Section 3.1). These are a class of grid-world problems, where popular exploration strategies struggle to find the optimal policy efficiently. We then provide empirical evidence that curiosity is a key property that helps overcome the challenges in this problem and is inherently present in optimistic initialization-based exploration strategies. In essence, curiosity encourages the agent to visit unseen states and speed up learning in many cases.

We then present a novel exploration strategy, denoted COIN, for Q-learning algorithms [69, 70]. COIN performs continual optimistic initialization of the Q-values through scaled optimistic initializations at “appropriate” steps. COIN is designed to overcome a major shortcoming in standard optimistic initialization [40, 42, 61] while preserving its beneficial traits. To the best of our knowledge, COIN is the first general exploration strategy that satisfies properties P1, P2, and P4 while obviating P3. We demonstrate the implications of satisfying these properties using the designed bridge crossing problem. We then present two variants of COIN, (V1) vanilla that is applicable to tabular Q-learning and (V2) dual-COIN that is appropriate for Q-learning with approximation. Lastly, we provide an empirical study comparing COIN against popular existing exploration strategies on benchmark tasks. We demonstrate that COIN is a general exploration strategy that scales to a range of tasks, both with dense and sparse reward functions. In particular, empirical results show that COIN performs consistently better or at par with existing strategies on domains with sparse feedback.

2 PRELIMINARIES

2.0 Markov decision processes: Stochastic control problems are commonly formulated as Markov decision processes (MDPs) [49] represented by $(S, A, P, R, \gamma)$, where $S$ is the state space, $A$ is the action space, $P : S \times A \times S \rightarrow [0, 1]$ is the transition function, $R : S \times A \rightarrow \mathbb{R}$ is the reward function, and $\gamma$ is the discount factor. An agent is assumed to follow an internal policy, $\pi : S \rightarrow A$, which maps states to actions. At each timestep, $t$, executing a visited action, $a_t$, at the current state, $s_t$, leads the agent to a new state, $s_{t+1}$, and results in a feedback signal, $r_t = R(s_t, a_t)$, representing the immediate utility gained from executing $a_t$ at $s_t$. A finite sequence of actions starting from an initial state $s_0$ results in an episode, $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_T)$. The cumulative discounted rewards thus obtained over $\tau$, also known as the episodic return, is defined as $R(\tau) = \sum_{i=0}^{T-1} \gamma^i R(s_i, a_i)$. The expected return for a policy $\pi$ is $J(\pi) = \mathbb{E}_{\tau \sim \pi} [R(\tau)]$. The optimal policy, $\pi^*$, is the policy that maximizes $J$, i.e., $\pi^* = \arg \max_\pi J(\pi)$.

2.0.2 Q-learning: The action-value function of $\pi$, denoted $Q^\pi(s_t, a_t)$, is the expected return over trajectories generated by executing $a_t$ at $s_t$ and then following $\pi$ from $s_{t+1}$ onward. The optimal Q-function, $Q^*(s_t, a_t)$, is concisely denoted by $Q^*(s_t, a_t)$. Off-policy Q-learning algorithms update $Q^*$ at each timestep via temporal difference (TD) learning [67, 69, 70], given by

$$Q^*_k(s_t, a_t) := Q^*_k(s_t, a_t) + \alpha_k \left( r_t + \gamma \max_{a_t'} Q^*_k(s_{t+1}, a_t') - Q^*_k(s_t, a_t) \right)$$

where $\alpha$ denotes the learning rate and $k$ denotes the Q-update iteration for $(s_t, a_t)$ pairs. The term, $r_t + \gamma \max_{a_t'} Q^*_k(s_{t+1}, a_t') - Q^*_k(s_t, a_t)$, is also known as the TD-error. Q-learning converges to $\pi^*$ with probability 1 provided that the following conditions hold: (1) the learning rate is well-behaved, i.e., $\sum_k \alpha_k = \infty$, $\sum_k \alpha_k^2 < \infty$, and (2) every state-action pair is visited infinitely often. A few curiosity-driven exploration [12, 17, 47] methods also fall under this category. While providing desired performance guarantees, such approaches come at the cost of additional memory requirements for storing state-action visitation counts [6, 7]. Alternatively, they are heavily dependent on the accuracy of an estimator approximating state-action visitation counts [28, 63].

2.0.3 Off-policy learning: RL algorithms are designed to optimize a target policy, $\pi$. Nonetheless, they might use a different policy, denoted $\eta$, to explore the environment. We refer to $\eta$ as the exploration strategy which is a mapping from states to actions. RL algorithms which provide policy convergence guarantees only for the case where $\eta \equiv \pi$ are known as on-policy algorithms [43, 54, 55], else allowing $\eta$ to differ from $\pi$, they are denoted off-policy [30, 38, 44]. Q-learning is a prominent example of an off-policy algorithm. While the target policy is defined by $\pi(s) := \arg \max_a Q(s, a)$, many cases the exploration strategy is different. For example, Q-learning is often paired with an $\epsilon$-greedy exploration strategy [60].

2.1 Related work

We follow the taxonomy presented by Amin et al. [4] to briefly discuss the major advantages and shortcomings of common exploration strategy classes for Q-learning.

1In general, $s_t$ might be sampled from an initial state distribution.
Table 1: Comparison of exploration strategies. We specify the category under which an exploration strategy ('Exp. strategy') falls as described in Section 2.1. We highlight which of the properties among coverage, infinite often visitation ('∞-often'), and curiosity is satisfied by each of the exploration strategies. We indicate whether each of them requires visitation tracking ('No vis. track'). 'Hyperparam.' is the hyperparameter essential to the exploration strategy.

<table>
<thead>
<tr>
<th>Exp. strategy</th>
<th>Category</th>
<th>Coverage</th>
<th>∞-often</th>
<th>No vis. track</th>
<th>Curiosity</th>
<th>Hyperparam.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε-greedy</td>
<td>C1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>ε</td>
</tr>
<tr>
<td>Boltzmann</td>
<td>C2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>Temperature</td>
</tr>
<tr>
<td>UCB</td>
<td>C2</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>Bonus coeff. (C_p)</td>
</tr>
<tr>
<td>Optimistic init.</td>
<td>C3</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>Initial Q-values</td>
</tr>
<tr>
<td>COIN</td>
<td>C3</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Bonus (b)</td>
</tr>
</tbody>
</table>

Definition 1 (P1: Coverage at limit). An exploration strategy is said to provide coverage at limit if \( \forall s \in S \times A, \prod_{i=0}^{\infty} (1 - P_r(s, a)) = 0 \).²

Any exploration strategy that does not provide coverage at the limit cannot guarantee convergence to the optimal policy as some actions along the optimal sequence of actions might never have been visited [61].

Definition 2 (P2: Infinite-often). Following the Borel-Cantelli lemma [19, 22], an \((s, a)\) pair is said to be visited infinitely often if \( \sum_{i=0}^{\infty} P_r(s, a) = \infty \). An exploration strategy is said to be 'Infinite-often' if every \((s, a)\) pair is visited infinitely often.

Note that an exploration strategy with the infinite-often property implies that it also has the property of coverage at the limit but not vice-versa. Various RL algorithms based on Watkins [70] and Singh et al. [56] guarantee convergence to the optimal policy if infinite-often visitation is satisfied [51]. Note that there is a difference between the convergence of \( \eta \) and \( \pi \). While \( \eta \) is not required to, and usually does not converge to \( \pi^* \), \( \pi \) can still converge to \( \pi^* \).

Definition 3 (P3: Visitation tracking). An exploration strategy using visitation tracking continually tracks (or estimates) the number of times each \((s, a)\) pair was visited.

See C2 in Section 2.1.2 for examples of exploration strategies that track the number of \((s, a)\) visits either explicitly in tabular MDPs or implicitly using an estimator in continuous state spaces.

Definition 4 (P4: Curiosity). An exploration strategy possesses curiosity if action visitation precedence at a given state is inversely correlated, in expectation, with the number of times that action was visited. That is, for any state \( s \) and action \( a \), the probability (sampling-based exploration) or score (hardmax exploration) for visiting \( a \) at \( s \) monotonically decreases, in expectation, with the number of visits to \( a \). In general, this may also occur after a bounded number of initial visits to \( a \) at \( s \).

Exploration strategies possessing the property of curiosity have been shown to outperform vanilla sampling-based methods on many hard-exploration tasks [18]. However, previous work provide vague definitions for curiosity. For example, the seminal curiosity work [47] defines this property as “intrinsic motivation/reward”. Such reward can be grouped into two broad classes: 1) encourage the agent to explore ‘novel’ states [12, 41, 48] or, 2) encourage the agent to perform actions that reduce the error/uncertainty in the agent’s ability to predict the consequence of its own actions (i.e. its knowledge about the environment) [20, 31, 45, 52, 57]. As a result, we provide a precise definition of curiosity (P4) that is general (applying to known curiosity-based algorithms) while being unambiguous. In common sampling-based exploration strategies (ε-greedy and Boltzmann), Q-values for specific \((s, a)\) pairs may increase over visitations. This would lead to a violation of P4. On the other hand, exploration in UCB-based algorithms, for instance, UCB1 [6] defines \( \eta(s) = \arg \max_a Q(s, a) + C_p \sqrt{\frac{\ln n}{N(s,a)}} \), where \( n \) is the total number of learning timesteps taken by the RL agent and \( N(s, a) \) is the number of times an \((s, a)\) pair has been visited. If \( C_p \) is set to a large enough value, the score for visiting \( a \) at \( s \) monotonically decreases with the number of visits to \( a \), hence satisfying P4. In optimistic initialization, with sufficiently large initialization of Q-values, the expected TD-error is non-positive.³ Since action visitation in optimistic initialization relies only on a hardmax over the Q-values, \( Q(s, a) \) for an action \( a \) visited at \( s \) with a small \( \alpha \), in expectation, decreases monotonically following Equation 1. Hence, optimistic initialization satisfies P4.

Table 1 presents a summary of exploration strategies and their associated properties.

3.1 Motivating example: the bridge problem

We further demonstrate the significance of curiosity on a toy example containing a high proportion of terminal states with undesirable outcomes. Consider the gridworld shown in Figure 1, which we refer to as the bridge crossing problem. We model it as an MDP. In the stationary bridge problem illustrated in Figure 1a, there are two terminating goals; one optimal (green cell at the right extreme) and another suboptimal (orange cell at the left extreme) along with many terminating negative outcomes for falling off the bridge (red cell). Assuming that the start state is fixed, an issue in this domain is that it is easy for the agent to converge on a suboptimal policy leading to the suboptimal goal, or take a long time to learn the optimal policy. This issue is particularly exacerbated when the horizon of the bridge, \( H \), is increased and sampling-based exploration strategies are deployed.

²See Even-Dar and Mansour [26] for the sufficiently large Q-initialization required to guarantee convergence.

³See Even-Dar and Mansour [26] for the sufficiently large Q-initialization required to guarantee convergence.
We demonstrate the shortcomings of popular exploration strategies listed in Table 1 using the bridge problem and explain how optimistic initialization addresses them.

3.1.1 Sampling-based: In this family of exploration strategies (Section 2.1.1) the probability of the agent following the greedy action over a sequence of states diminishes exponentially as $T$ increases. Reaching the optimal goal in this example clearly requires $T \geq H - \lfloor (H - 1)/2 \rfloor$ steps. The probability of the agent reaching the optimal goal state by following the optimal greedy policy in an episode is $\lim_{T \to \infty} \prod_{t=1}^{T} Pr(a_t|s_t, \pi^*) = \lim_{T \to \infty} (1 - \epsilon)^T = 0$, where $\epsilon$ is the non-zero probability of selecting a non-greedy action at each step. That is, the probability of reaching the optimal goal state, diminishes exponentially with the bridge length even when $\pi^*$ is known. This results in sampling-based exploration strategies becoming sample-inefficient in similar settings [27, 58].

3.1.2 Optimistic initialization-based: Since optimistic initialization possesses curiosity, it encourages the RL agent to explore unvisited $(s, a)$ pairs. This, in practice, results in optimistic initialization being sample-efficient on the bridge problems.

3.1.3 Limitations of optimistic initialization: Optimistic initialization is known to potentially converge to suboptimal solutions in stochastic environments since it does not hold the infinite-often property [26, 27]. In such scenarios, an appropriate $Q$-initialization may be a tedious hyperparameter to tune. Furthermore, optimistic initialization has limited efficacy in non-stationary MDPs as exploration ceases when $Q$-values converge.

4 CONTINUOUS OPTIMISTIC INITIALIZATION

We turn to present our proposed continual optimistic initialization exploration strategy, denoted $COIN$, an optimistic initialization-based exploration strategy, to address the shortcomings of “vanilla” optimistic initialization. As such, $COIN$ follows $\pi(s) = \arg \max_a Q(s, a)$ and periodically augmenting the $Q$-values in a way that achieves effective exploration.

Based in $Q$-learning, $COIN$ guarantees that the $Q$-values for each $(s, a)$ pair along all the greedy trajectories will converge after a finite number of $Q$-value updates [69]. We define a greedy trajectory as a sequence of $(s, a)$ pairs obtained when following the greedy policy, i.e., $\pi^g(s) = \arg \max_a Q(s, a)$. Once $Q$-values converge along all greedy trajectories, we add a positive bonus, $b$, to the $Q$-function across all $(s, a) \in S \times A$. We term the step at which such convergence occurs as an “appropriate $b$ addition step”. Intuitively, adding $b$ is analogous to initializing the $Q$-function with optimistic values, albeit continually. A pseudocode of $COIN$ has been provided in Algorithm 1 respectively.

Definition 5 (Appropriate $b$ addition step). We define an appropriate $b$ addition step as the training step at which the maximum TD-error along the greedy trajectories is within an $\epsilon$ threshold.

We also call an appropriate $b$ addition as a $COIN$ iteration to reduce verbosity. Note there may be multiple episodes between two consecutive $COIN$ iterations.

4.0.1 $b$ setting in $COIN$. Although any positive $b$ is sufficient for the theoretical properties of $COIN$ to hold, empirically we observe that setting $b$ such that

$$b = \max_{s \in S^T} \left\{ \max_{(s,a) \in A} Q(s,a) - \min_{(s,a) \in A} Q(s,a) \right\} + \epsilon_b,$$
where $S^0$ is the set of states reachable under $\eta$ and $\epsilon_b$ is a small positive value, results in sample-efficient learning. This heuristic makes all actions along the greedy trajectory to be optimistic with respect to the current $\pi^\eta$.

More formally, COIN follows: every COIN iteration, set $\forall (s, a) \in S \times A, Q(s, a) = Q(s, a) + b$. Adding $b$ to the $Q$-function continually at an infinite number of COIN iterations ensures infinite-often visitation (P2) (see Lemma 1).

4.1 Theoretical properties of COIN

The following theoretical properties and analysis apply to tabular $Q$-learning. While COIN can be coupled with function approximation as we show later in section 5.2, our theoretical claims might no longer be guaranteed.

**Lemma 1 (COIN ensures infinite often visitation (P2)).** Assuming that the $Q$-function is updated at each COIN iteration, any $(s, a) \in S \times A$ must be visited infinitely often under COIN, i.e., $\Sigma_{t=0}^\infty Pr_t(s, a) = \infty$.

**Proof.** We prove via induction that any $(s, a)$ pair must be visited within a bounded number of COIN iterations. As a result, an infinite number of iterations would result in infinite-often visitation.

**Base case:** (every action, $a \in A$, will be visited at $s$ after a bounded number of visitations to $s$, and specifically for $s = s_0$.) By contradiction, we assume some action, $a'$, at $s$ will not be visited within a finite number of COIN iterations.

**Notation:**
- $Q^{(old)}(s, a)$: Q-function at the end of the COIN iteration when $a'$ was last visited at $s$ or iteration $0$ if it was never visited.
- $Q^{(m)}$: Q-function after $m$ COIN iterations since $a'$ was last visited.
- $\alpha$: $\text{arg max}_a Q^{(m)}(s, a)$.

Since $a'$ hasn’t been visited by the greedy policy in $m$ COIN iterations,
\[
Q^{(m)}(s_0, a') > Q^{(m)}(s_0, a^*), \forall m
\]
\[
Q^{(m)}(s_0, a') \geq Q^{(old)}(s_0, a') + m\epsilon_b
\]
As $Q^{(m)}$ is updated following Equation 1, for a low enough $\alpha$,
\[
Q^{(m)}(s_0, a^*) \leq Q^*(s_0, a^*) \leq C
\]

Since $R$ is bounded, $C$ is finite. From Equations (2) and (3) it must be that $\exists m$ such that,
\[
Q^{(old)}(s, a') + m\epsilon_b > C
\]

Equations (2) and (3), in conjunction, contradict (5) as they imply $C < C$. Hence, $a'$ must be visited within a finite number COIN iterations. Further, since $b > 0$, an infinite number of COIN iterations will result in $a'$ being visited infinitely often.

**Induction assumption:** (every action, $a \in A$, at $s_{n-1}$ will be visited after a bounded number of visitations to $s_{n-1}$, where $s_{n-1}$ is any state reachable from $s_0$ in $n$ steps.)

**Induction step:** Following the (general claim) base case, every action at $s_{n-1}$ will be visited infinitely often. Hence, for any state, $s_n$, if $Pr(s_n | s_{n-1}, a) > 0$, it is reachable in $(n+1)$ steps from $s_0$ with a non-zero probability. As such, the general base case can be invoked again. That is, any $a'$ will be visited at $s_n$ an infinite number of times.

\[\Box\]

**Corollary 1 (COIN satisfies coverage at limit (P1)).** Assuming that the $Q$-function is updated at each COIN iteration, any $(s, a) \in S \times A$ must be selected under COIN, i.e., $\prod_{t=0}^\infty (1 - Pr_t(s, a)) = 0$.

Corollary 1 follows from Lemma 1 since infinite-often visitation implies coverage at limit.

**Remark 1.** COIN has the property of curiosity.

After a finite number of COIN iterations, $Q(s, a)$ for any unvisited $(s, a)$ pair must be overestimated as its $Q$-values are continually inflated. For a small enough $\alpha$, $Q(s, a)$ for visited $(s, a)$ pairs are also overestimated after a finite number of COIN iterations as any reduction in their $Q$-values will be overcome by the addition of $b$ to the $Q$-function. Once, the $Q$-values for all $(s, a)$ pairs are overestimated, they are updated following Equation 1 with a negative TD-error, in expectation. This leads to a monotonic decrease in the $Q$-values, in expectation, with each visitation to $a$.

4.2 Extension of COIN to non-stationary MDPs

COIN can be interpreted as a model-free restarting strategy for non-stationary MDPs (if $b$ were held fixed and equivalent to optimistic initialization). Such strategies have been suggested previously for non-stationary episodic MDPs and multi-armed bandit problems [3, 14]. COIN may be a suitable candidate for non-stationary MDPs since it has infinite-often property. However, convergence of RL algorithms in non-stationary MDPs is an active branch of research as optimal convergence proofs of RL algorithms often assume stationary transition and reward functions [46]. Notably, prior work has shown value-function-based RL to produce policies "close" to optimal in a special class of non-stationary MDPs [23]. Section 4.3.2 briefly discusses the effectiveness of COIN on a toy non-stationary MDP. However, an in-depth analysis is left for future work.

4.3 COIN on the bridge problem

We provide empirical results for COIN and the exploration strategies presented in Table 1 when coupled with tabular $Q$-learning on a few variants of the bridge problem. Later in Section 5.2, we present results on domains with continuous state spaces.

Since COIN has a deterministic policy, it avoids issues faced by sampling-based exploration strategies. COIN’s infinite-often visitation property overcomes the limitation of optimistic initialization. Note that it is expected to observe “dips” in the learning curves with COIN each time a $b$ addition is performed as it perturbs the $Q$-values and encourages the greedy policy to explore. In order to clearly present these trends, we perform a $b$ addition after a fixed number of episodes instead of our proposed $b$ addition step from Definition 5. $\epsilon_b$ and $\epsilon$ are set to 0.1 and 0.05 respectively.

4.3.1 Results for stationary bridge: Figures 2a and 2b show results for the case where $H=15$ with deterministic and stochastic transitions respectively. In the deterministic case, it can be observed that all strategies except $\epsilon$-greedy find the optimal solution within 6,000 episodes of learning. While in the stochastic case, Optimistic init. fails as learning progresses since it suffers from the issue explained
in Section 3.1.3 whereas COIN does not since it visits \((s,a)\) pairs infinitely often via continual \(b\) additions. The results in Figures 2c and 2d, where \(H=31\), offer similar insights but in addition, highlight the advantages of optimistic initialization (Figure 2c) and COIN. We speculate that the advantage stems from their curiosity property.

Note that we use a count-based intrinsic reward proposed in Bellemare et al. [12] and combine it with \(\epsilon\)-greedy exploration (Curious \(\epsilon\)-greedy). This approach leads to improved learning efficiency over \(\epsilon\)-greedy demonstrating that curiosity is helpful in this domain.

**4.3.2 Results for non-stationary bridge.** Although optimistic initialization with \(Q\)-learning can converge to an optimal policy, it is known to potentially reach a local optimum when the MDP is non-stationary. Extending the bridge problem, let a new optimal goal emerge at a new location. At the same time, a new path is added such that it leads to the new optimal solution from the start state. An illustration of the new problem is shown in Figure 1b. We allow \(Q\)-learning to first converge on the old setting and require it to adjust to the new setting. The results in Figures 2e and 2f on the non-stationary bridge (\(H=7\)) tasks show that all methods except COIN fail to reach the optimal solution. Since optimistic initialization does not visit state-action pairs infinitely often, it fails to find the new optimal solution whereas COIN succeeds in doing so.

These results positively answer empirical questions on the capability of COIN in solving a simple instance of a non-stationary MDP by updating \(b\) continually, demonstrating its ability to successfully address a major drawback of “vanilla” optimistic initialization.

**5 DUAL-COIN**

We present dual-COIN, an alternative view of COIN, where, instead of adding \(b\) to the \(Q\)-function, we modify the reward function to induce a similar effect. Consider the case where for \(b > 0\) the reward function is updated such that \(R_{b} := R - b\). That is we subtract a positive constant from the reward function. It can be shown that the \(Q\)-function learned using \(Q\)-learning converges to \(Q_{b} = Q - \frac{b}{1-\gamma}\), where \(Q\) is the \(Q\)-function learned on \(R\) (see Lemma 2 for proof). A similar observation was also presented in prior work [42, 58] for the optimistic initialization case.

Note that, when the MDP has a large or continuous state space, a function approximator is commonly utilized to represent the \(Q\)-function. Updating the \(Q\)-function to \(Q(s,a) := Q(s,a) + \beta\) for all \((s,a) \in S \times A\) might become intractable in this case. Dual-COIN makes COIN practical in such settings.
5.1 Theoretical properties of dual-COIN

As stated in section 4.1, we remind the reader that the following theoretical properties and analysis apply to tabular Q-learning.

**Lemma 2 (Additive property of Q-values).** If a constant, b, is uniformly added to R, Q-values updated using Q-learning converge to $Q + \frac{b}{1 - \gamma}$ for an infinite-horizon MDP.

**Proof.**

\[
Q(s_t, a_t) = \mathbb{E} \left[ r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \right]
\]

Let $R(s, a) := R(s, a) + b, \forall (s, a) \in S \times A$ and $Q_{+b}$ be the new Q-function.

\[
Q_{+b}(s_t, a_t) = \mathbb{E} \left[ r_t + b + \gamma \max_{a_{t+1}} Q_{+b}(s_{t+1}, a_{t+1}) \right] = \mathbb{E} \left[ r_t + b + \gamma \left( r_{t+1} + b + \gamma \max_{a_{t+2}} Q_{+b}(s_{t+2}, a_{t+2}) \right) \right]
\]

\[
= b \sum_{t=1}^{T} y^t + \mathbb{E} \left[ r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \right] = b \frac{1 - y^T}{1 - y} + Q(s_t, a_t)
\]

For an infinite horizon MDP, i.e., $T \to \infty$,

\[
Q_{+b}(s_t, a_t) = Q(s_t, a_t) + \frac{b}{1 - \gamma}
\]

\[\square\]

5.1.1 b setting in dual-COIN:

\[
b_{dual} = \left( \frac{1 - y}{1 - y^T} \right) \max_{s \in S^f} \left( \max_{(s, a) \in A} Q(s, a) - \min_{(s, a) \in A} Q(s, a) \right) + \epsilon_b,
\]

where $S^f$ is the set of the states visited under $\eta$ so far and $\epsilon_b$ is a small positive value.

**Lemma 3 (Dual-COIN ensures infinite often visitation (P2)).** Assuming that $R$ is updated at each COIN iteration, any $(s, a) \in S \times A$ must be visited infinitely often under dual-COIN, i.e., $\sum_{n=0}^{\infty} P_T(s, a) = \infty$.

**Proof.** We prove via induction that any $(s, a)$ pair must be visited within a bounded number of COIN iterations. As a result, an infinite number of iterations would result in infinite-often visitation.

**Base case:** (every action, $a \in A$, will be visited at least after a bounded number of visitations to $s$).

**Notation:**

- $Q^{(m)}$: Q-function after $m$ COIN iterations since $a'$ was last visited or never visited.
- $Q^{(m)}$: arg max$_a Q^{(m)}(s, a)$.

Let $a'$ be unvisited by the greedy policy in $m$ COIN iterations,

\[
Q^{(m)}(s_0, a') < \max_a Q^{(m)}(s_0, a), \forall m
\]

As $Q^{(m)}(s_0, a')$ is updated following Equation 1, from Lemma 2, for a low enough $\alpha$,

\[
\max_a Q^{(m)}(s_0, a) < Q^* \left( \arg \max_a Q^{(m)}(s_0, a) \right) - \alpha \epsilon_b \left( 1 - \frac{1}{1 - \gamma} \right)
\]

Since, $R$ is bounded, $Q^*$ is bounded. From Equations (6) and (7) it must be that $\exists m$ such that,

\[
Q^{(m)}(s_0, a') = \max_a Q^{(m)}(s_0, a)
\]

Hence, $a'$ must be visited within a finite number of dual-COIN iterations. Further, since $b_{dual} > 0$, an infinite number of COIN iterations will occur resulting in $a'$ being visited infinitely often.

**Induction assumption:** (every action, $a \in A$, at $s_{n-1}$ will be visited after a bounded number of visitations to $s_{n-1}$, where $s_{n-1}$ is any state reachable from $s_0$ in $n$ steps.)

**Induction step:** Following the (general claim) base case, every action at $s_{n-1}$ will be visited after a bounded number of visitations to $s_{n-1}$. Hence, for any state, $s_0$, if $P_T(s_0, s_{n-1}) > 0$, it is reachable in $(n + 1)$ steps from $s_0$ with a non-zero probability. As such, the same argument as the one presented in the base case can be applied. Thus, $a'$ will be visited at $s_n$ an infinite number of times. \[\square\]

**Corollary 2 (Dual-COIN satisfies coverage at limit (P1)).** Assuming that $R$ is updated at each COIN iteration, any $(s, a) \in S \times A$ must be selected under dual-COIN, i.e., $\prod_{n=0}^{\infty} (1 - P_T(s, a)) = 0$.

Corollary 2 follows from Lemma 3 since infinite often visitation implies coverage at the limit.

**Remark 2.** Dual-COIN has the property of curiosity.

After a finite number of COIN iterations, subtracting $b$ from $R$ continually leads to underestimated Q-values along the greedy trajectories. Specifically, as Q-values of visited $(s, a)$ pairs decrease, the expected TD-error is negative. Since action visitation in COIN relies only on a hardmax over the Q-values, similar to optimistic initialization, $Q(s, a)$ for an action $a$ visited at $s$ with a small $\alpha$, in expectation, decreases monotonically.

5.2 Empirical study

5.2.1 Domains: We evaluate dual-COIN on 6 benchmark domains with continuous state spaces, from the OpenAI gym [16] and MiniGrid [21], covering both dense and sparse reward functions. The domains with sparse reward functions were also used in Sun et al. [58] to demonstrate the effectiveness of their proposed optimistic initialization-based approach on tasks with sparse rewards. In 3 of the 4 OpenAI gym domains, i.e., ‘Cartpole-v1’, ‘Acrobot-v1’, and ‘LunarLander-v2’ have dense reward functions. The 4th domain, ‘MountainCar-v0’, has a sparse reward function. A positive reward is received by the agent only upon reaching the goal state, and a negative reward at all other states. In the MiniGrid domains, i.e., ‘Empty-Random-6x6-v0’ and ‘MultiRoom-N2-S4-v0’ the state is only partially observable (agent-local view of the grid only). MiniGrid domains also have a sparse reward function where a positive reward is given only on reaching the goal state and 0 otherwise.

5.2.2 Baselines. We use exploration strategies belonging to categories C1 (sampling-based), i.e., $\varepsilon$-greedy, and Boltzmann, and C3
(optimism-based), i.e., Optimistic (norm.) [42], as baselines for comparison since, similar to COIN, they do not require learning any additional estimator. We employ vanilla deep Q-network learning (DQN) [44] adapted from Achiam [1] as the underlying learning algorithm for all of the baselines. These experiments aim to demonstrate that COIN is a general exploration strategy and can perform competitively with respect to popular exploration strategies without additional assumptions.

5.2.3 Results. The graphs presented in Figure 3 provide a positive answer to the generalizability and competitiveness of dual-COIN. Dual-COIN outperforms the baselines in terms of sample efficiency on 4 out of 6 domains, namely, 'CartPole-v1' (Figure 3a), 'LunarLander-v2' (Figure 3c), 'MountainCar-v0' (Figure 3d), and 'MultiRoom-N2-S4-v0' (Figure 3e) while performing on par in the remaining domains. In particular, we notice that in the sparse reward tasks, dual-COIN consistently has better sample efficiency. We believe that dual-COIN's property to induce curiosity plays a major role in this. Complete details of the domains, hyperparameter settings of dual-COIN and the baselines, and the codebase for these experiments are available at https://github.com/Pi-Star-Lab/coin.

6 SUMMARY
We present a novel optimistic initialization-based approach, COIN, possessing a unique set of properties associated with effective exploration strategies. It performs continual optimistic initialization of Q-values to overcome the limitations of optimistic initialization in stochastic and non-stationary environments. We provide theoretical evidence of COIN possessing infinite-often visitation property which helps it overcome these limitations. We validate our claims on the bridge crossing problem. Compared to common existing exploration strategies, we demonstrate the superiority of COIN in solving long-horizon stochastic and non-stationary bridge problems. Extending COIN to continuous state spaces, we then present dual-COIN. Empirical results on 6 benchmark domains support our claim that COIN is a general exploration strategy by outperforming 3 common existing exploration strategies on 4 out of 6 domains. We observe that COIN is more effective than these strategies in sparse reward benchmark domains which we speculate is a result of its curiosity-driven behavior.

ACKNOWLEDGMENTS
The reported work has taken place in the PiStar AI and Optimization Lab at Texas A&M University. PiStar is supported in part by NSF (IIS-2238979).


