Unpacking Reward Shaping: Understanding the Benefits of Reward Engineering on Sample Complexity

Abhishek Gupta*
University of Washington
abhgupta@cs.washington.edu

Aldo Pacchiano*
Microsoft Research, NYC
apacchiano@microsoft.com

Yuexiang Zhai
UC Berkeley, EECS
simonzhai@berkeley.edu

Sham M. Kakade
Harvard University
sham@seas.harvard.edu

Sergey Levine
UC Berkeley, EECS
svlevine@eecs.berkeley.edu

Abstract

Reinforcement learning provides an automated framework for learning behaviors from high-level reward specifications, but in practice the choice of reward function can be crucial for good results – while in principle the reward only needs to specify what the task is, in reality practitioners often need to design more detailed rewards that provide the agent with some hints about how the task should be completed. The idea of this type of "reward-shaping" has been often discussed in the literature, and is often a critical part of practical applications, but there is relatively little formal characterization of how the choice of reward shaping can yield benefits in sample complexity. In this work, we build on the framework of novelty-based exploration to provide a simple scheme for incorporating shaped rewards into RL along with an analysis tool to show that particular choices of reward shaping provably improve sample efficiency. We characterize the class of problems where these gains are expected to be significant and show how this can be connected to practical algorithms in the literature. We confirm that these results hold in practice in an experimental evaluation, providing an insight into the mechanisms through which reward shaping can significantly improve the complexity of reinforcement learning while retaining asymptotic performance.

1 Introduction

Reinforcement learning (RL) in its most general form presents a very difficult optimization problem: when there are no constraints on the reward function or dynamics, a learning algorithm may need to exhaustively explore the entire state space to discover high-reward regions. Naïve algorithms that rely entirely on random exploration are known to be exponentially expensive [26, 3], and much of the theoretical work on efficient RL algorithms has focused on smarter exploration strategies that aim to more efficiently cover state space, typically in time polynomial in the state space cardinality [6, 8, 24, 55, 27, 32, 52]. Much of this work is based on upper confidence bound (UCB) principles and prescribes some kind of exploration bonus to prioritize exploration of rarely visited regions. Analogous strategies have also been employed in a number of practical RL algorithms [41, 40, 10,

[*] Equal Contributions

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

12, 22, 36, 43, 42, 37]. However, perhaps surprisingly, much of the empirical work on reinforcement learning does not make use of explicit exploration bonuses or other dedicated exploration strategies, despite numerous theoretical results showing them to be essential to attain tractable sample complexity. Instead, practitioners often incorporate prior knowledge of each task into designing or shaping the reward function [31, 38, 39, 44, 5, 11, 50, 54], preferring this heuristic approach over the more principled exploration strategies. At this point, one may wonder why is reward shaping often practically preferable to dedicated exploration?

A likely answer to this question lies in the fact that even the best general-purpose exploration strategies still require visiting every state in the MDP at least once in the worst case. This is of course unavoidable without further assumptions. However, in practice, sample complexity that is polynomial in the size of the entire state space might *still* not be practical, and hence prior knowledge in the form of reward shaping is required to render such tasks tractable. Surprisingly, despite the widespread popularity of reward shaping in RL applications, the analysis of reward shaping has remained limited to proving policy invariance [31] or largely empirical observations, often relegating reward shaping to folk knowledge. In this work, we take a step towards studying the effect of reward shaping on the efficiency of RL algorithms, by asking the following question:

Can we theoretically justify the sample complexity benefits that reward shaping from prior domain knowledge can provide for reinforcement learning?

We aim to provide a set of tools that formally analyze how reward shaping can improve the complexity of tabula-rasa RL and better direct exploration. To perform this analysis, we first propose a simple modification to standard RL algorithms —"UCBVI-Shaped" that incorporates shaped rewards into optimism based exploration. We use this algorithm instantiation to then provide a regret analysis framework that studies how this introduction of shaped rewards can (in certain cases) provide much more directed optimism than uninformed exploration, while maintaining asymptotic performance.

To approach our analysis, we specifically consider problems where the reward shaping is provided through a term $\widetilde{V}$, a (potentially suboptimal) approximation of the optimal value function $V^\star$. In particular, we assume that the shaping $\widetilde{V}$ is a multiplicatively bounded approximation of the optimal value function $V^\star$, i.e., $\widetilde{V}(s) \leq V^\star(s) \leq \beta\widetilde{V}(s), \forall s$, for a finite multiplicative factor β. This type of shaped reward function $\widetilde{V}$ can be incorporated into a standard RL algorithm like UCBVI [8] through two channels: (1) bonus scaling – simply reweighting a standard, decaying count-based bonus $\frac{1}{\sqrt{N_h(s,a)}}$ by the per-state reward shaping and (2) value projection – adaptively projecting learned value functions into ranges of value functions derived from the reward shaping.

We show that this relatively simple algorithmic instantiation lends itself to an analysis that shows significant sample complexity benefits with shaping. Intuitively, the key pieces in our complexity analysis of UCBVI-Shaped are: (1) A multiplicative sandwich condition (via β) between $\widetilde{V}$ and $V^\star$ allows for the bonus scaling to depend on $\beta\widetilde{V}$ instead of a coarse approximation of $V^\star$ such as H. This allows for a reduction of complexity from a horizon H dependent term to one scaling with $\beta\widetilde{V} \leq \beta V^\star$ by allowing for a faster decay of the exploration bonuses while still providing enough optimism. (2) A projection of the value function prevents "over optimism" by hastening the convergence of the empirical $\widehat{Q}$ functions during value iteration, thus allowing for faster detection of sub-optimal actions. This results in the ability to prune out large parts of the state space, as we also validate empirically.

To summarize, the key contribution of this work is to characterize how reward shaping can provably improve sample efficiency by providing gains in both $|\mathcal{S}|$ and H dependent terms. We do so by analyzing the gains from reward shaping in two different ways: bonus scaling and value projection. We show that the "quality" (determined by β) of the provided shaping can significantly improve the sample efficiency of the resulting reinforcement learning algorithm, provide a set of analysis techniques to understand improvements in sample complexity from shaping, and confirm our findings with numerical experiments.

2 Related Work

Regret Analysis in Finite Horizon Episodic Tabular MDPs. Recent research on regret analysis has studied both model-based and model-free RL methods. Model-based methods [23, 4, 8, 20, 17, 35] first learn a model from previous experience and use the learned model for future decision making. In contrast, model-free methods [24, 9, 55, 46, 30, 27] aim to learn the value function without the model estimation stage and use the learned value function for decision making. Our analysis lies in the model-based framework and is similar to the setting of [8] but with additional assumptions on knowing $\widetilde{V}$ as a multiplicatively bounded approximation of $V^\star$. The main difference between our results and the aforementioned works is that we consider the reward shaping setting with a truncated interval assumption. As a result, we reduce the state dependency $|\mathcal{S}|$ to the some smaller "effective state space". Our work is also closely related to [21]. We discuss this in Section 3.

Regret Analysis with Linear Function Approximation. A recent line of literature has investigated the linear function approximation setting by assuming the transition kernels or the value functions can be represented by d-dimensional linear features [25, 53, 7, 56, 13, 48, 47, 45] or even general function classes [49]. With the aforementioned assumptions, the regret analysis will only depend on the ambient dimension d (or other intrinsic complexity measure), instead of S, A in the tabular setting [45, 48, 13, 56, 7, 52, 25], which could greatly decrease the complexity of learning. Instead of posing structural assumptions on the function class representing the MDP's values or dynamics, we ask the question of whether having approximate knowledge of $V^\star$ can improve the speed of learning.

Practical Reward Design and Reward Shaping. Ng et al. [31] proposed a potential-based shaping function F that ensures policy invariance under transformation. However, unbiased potential based reward shaping is rarely used in practice. In most applications, heuristic reward design is carefully performed with potentially biased reward functions. In some large-scale practical RL tasks [11, 50, 51], the reward functions are heavily handcrafted based on prior knowledge. Besides reward engineering, a distinct line of work applies uninformed exploration algorithms like count-based reward shaping or intrinsic rewards to encourage greater state visitation [41, 40, 10, 12, 22, 36, 43, 42, 37, 28]. Importantly, these methods optimize for the worst case, as they try to cover *all* states since the exact location of the reward is unknown. In contrast, we look at the problem of incorporating domain knowledge via reward shaping into the exploration process. Closely related to our work is that of Cheng et al. [14], which studies how to incorporate shaping (heuristics) into the RL process via reducing the effective horizon. This work provides gains by reducing the horizon factor while our work provides gains by reducing the size of the effective state space that needs to be searched through.

3 Overview

We consider an episodic Markov Decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}^\star, r, H)$ where $\mathcal{S}$ corresponds to the state space, $\mathcal{A}$ is the action space, $\mathbb{P}^\star$ is the transition operator, $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function and H is the problem horizon. We use $|\mathcal{S}|, |\mathcal{A}|$ to denote the number of state and actions, and we use $\mathbb{P}^\star(\cdot|s,a)$ to denote the transition probability of state action pair (s,a). The value function $V^\pi(s_0)$ of a policy π starting at an initial state s_0 is defined as $V^\pi(s_0) := \mathbb{E}_\pi\left[\sum_{h=0}^{H} r(s_h, a_h)\right]$, where $\mathbb{E}_\pi$ denote the transition dynamic of $\mathcal{M}$ under π. Similarly, $V^\star(s_0)$ is the value function of the optimal policy $\pi^\star$. We consider a sequential interaction between a learner and the MDP $\mathcal{M}$ occurring in rounds indexed by $t \in \mathbb{N}$. At the start of round t the learner selects a policy π_t that is used to gather a sample trajectory from $\mathcal{M}$. As is standard in the literature, we measure the learner's performance up to round T by $\text{Regret}(T) := \sum_{t=1}^{T} V^\star(s_0) - V^{\pi_t}(s_0)$.

Our goal is to show that knowing a reward shaping term $\widetilde{V}$ allows for significantly more sample efficient learning, which with high probability has a sublinear regret upper bound. This bound has a leading term that depends on an "effective state space" of $\mathcal{M}$, determined by the quality of a reward shaping term $\widetilde{V}$ and the nature of the particular MDP being solved. This pruned effective state space can be much smaller than $|\mathcal{S}|$. Additionally, we will show an improved horizon dependence as the shaping term $\widetilde{V}$ allows for the bonus terms to be smaller and therefore decay faster thus replacing horizon factors of H with $\beta\widetilde{V}$ ones.

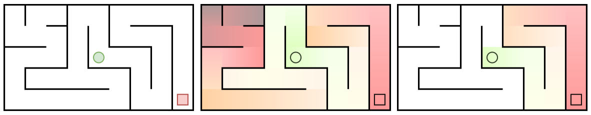


Figure 1: Reward shaping can allow for reduction of the "effective" state space size. In the above maze environment, the initial state is in the middle of the maze so finding the goal only requires solving half of the maze **(Left)**. While uninformed state covering exploration **(Middle)** would go both directions since it has no knowledge of goal location, effective shaping **(Right)** would allow for halving of the effective state space. The green circle represents the starting point and the red square represents the goal. Different colors represents the landscape of the shaped value function (green indicates smaller value and red indicates larger value).

Intuitively, reward shaping allows for the consideration of a reduced effective state space. Consider the maze environment in Fig. 1 with agent starting in the middle. Without knowing where the goal is, an uninformed exploration algorithm needs to explore nearly the entire maze. However, an effectively incorporated reward shaping term $\widetilde{V}$ would allow the algorithm to prune out the entire left half of the maze, effectively halving the effective state space size. This can happen in two ways: firstly, if the shaping can directly indicate that certain actions are sub-optimal, then entire regions of the space can be eliminated from search. Secondly, even if states are not eliminated, if their corresponding bonuses are scaled according to their shaped value $\widetilde{V}$ from the shaping rather than uniformly with H, it limits unnecessary exploration of suboptimal states.

Based on this intuition, we propose modifications to a standard reinforcement algorithm that allows us to perform this analysis effectively. To aid our analysis, we introduce UCBVI-Shaped, a modification of the UCBVI algorithm [8] that uses an additional reward shaping term $\widetilde{V}$ in two ways: (1) Bonus scaling: $\widetilde{V}$ is used to provide an exploration bonus that combines inverse state visitation counts with reward shaping. This allows for the reduction of overoptimism and the dropping of the horizon H dependence (2) Value projection: the shaping term is used to clip the empirical value function $\widehat{V}_h(s)$, which enables pruning unnecessary elements of the state space and allows for a bound that depends only on the effective state space size. Informally, our main result for UCBVI-Shaped can be summarized as follows:

Theorem 3.1 (Main Informal)**.** *With probability at least $1-\delta$, the regret of UCBVI - Shaped satisfies*[2]

$$\mathrm{Regret}(T) = \mathcal{O}\left(B(\widetilde{V},\mathcal{M})\log(T/\delta) + \mathrm{Regret}-\mathrm{UCBVI}(\mathcal{S}_{\mathrm{remain}},\mathcal{A},T)\right).$$

Here, $\mathcal{S}_{\mathrm{remain}}$ corresponds to the set of states in $\mathcal{M}$ where the information contained in $\widetilde{V}$ was not enough to rule them out, and $\mathrm{Regret}-\mathrm{UCBVI}(\mathcal{S}_{\mathrm{remain}}\mathcal{A},T)$ is a UCBVI regret upper bound function evaluated on $\mathcal{S}_{\mathrm{remain}}$ and $\mathcal{A}$. $B(\widetilde{V},\mathcal{M})$ is a time-agnostic complexity measure that depends solely on the $\mathcal{M}$ and the quality of the shaping $\widetilde{V}$.

We will define $\mathcal{S}_{\mathrm{remain}}$ and $B(\widetilde{V},\mathcal{M})$ and describe their relationship to $\widetilde{V}$ precisely in Section 5. If $\mathcal{S}_{\mathrm{remain}} \ll \mathcal{S}$, the regret of UCBVI - Shaped can be substantially smaller than $\mathrm{Regret}-\mathrm{UCBVI}(\mathcal{S},\mathcal{A})$, particularly since the first term of the regret grows logarithmically while the second scales with $\sqrt{T}$. We state our assumptions next:

Assumption 1. *The quality of the shaping term $\widetilde{V}$ is described by a parameter β. We assume access to a shaping "value function" estimators $\widetilde{V}_h : \mathcal{S} \to \mathbb{R}$ such that $V_h^\star(s) \leq \beta \widetilde{V}_h(s)$, for all $s \in \mathcal{S}$, $h \in [H]$ and for some $\beta \geq 1$.*

Instead of absorbing β into the definition of $\widetilde{V}_h$ we allow $\widetilde{V}_h(s) < V_h^\star(s)$ for some states $s \in \mathcal{S}$. We'll show that learning a value of β that turns this assumption true can be performed online. This assumption is intimately related to the optimistic $\widetilde{Q}$ assumptions of [21]. A thorough comparison with this work can be found in Appendix B.2.

Assumption 2. *We assume the reward functions satisfy $r(s,a) \in [0,1], \forall (s,a) \in \mathcal{S}\times\mathcal{A}$.*

Assumption 3. *The states $\mathcal{S}$ are $h-$indexed, i.e., the states reachable at time $h \in [H]$ are disjoint from the states reachable at time $h' \in [H]$ when $h \neq h'$.*

Contributions. Our main conceptual innovation is to introduce the notions of pseudosuboptimal and path-pseudosuboptimal states to quantify the "effective" size of the state space as a function of the quality of the shaping term $\widetilde{V}$ and use these notions to show how UCBVI-Shaped can attain significantly improved regret rates. In contrast with for example [21], where the regret rates will always scale at least with the number of states, our regret guarantees depend on an effective state size that may be orders of magnitude smaller. We believe this captures the real complexity improvement that reward shaping may yield, namely, avoid exploration of unnecessary areas of the state space. Other approaches such as [21] may (in general) at most yield an improvement in the dependence on the effective size of the action space. We further show that incorporating shaping into the exploration bonus term improves the horizon-dependence in the bound when the shaping is good enough, allowing us to replace a leading H term with $\max_s \beta\widetilde{V}(s)$.

[2] Our main result, Theorem 5.2, is slightly more complex than this statement. We have chosen this simplified form to aid the reader to form the right intuition.

4 The UCBVI-Shaped Algorithm

To perform analysis of reward shaping, we build on the framework of the UCBVI algorithm [8]. UCBVI is an exploration algorithm based on the upper confidence bound, described in detail in Appendix D. This forms a convenient base algorithm for incorporating shaped rewards in a way that admits faster learning while maintaining asymptotic performance (as we will discuss in Section 5). Our algorithm, UCBVI-Shaped, is a combination of two changes to the upper confidence bound algorithm. First, we modify the bonus scaling to depend on $\widetilde{V}$. Second, we introduce a projection subroutine into value iteration, implemented as a standard clipping operation.

Algorithm 1 UCBVI - Shaped

```
Input reward function $r$ (assumed to be known), confidence parameters
for $t = 1, \ldots, T$
    Compute $\widehat{P}_t$ using all previous empirical transition data as $\widehat{\mathbb{P}}_t(s'|s,a) := \frac{N_h^t(s,a,s')}{N_h^t(s,a)}$, $\forall h, s, a, s'$.
    Compute reward bonus $b_h^t(s,a)$ from Eqn. 1 (roughly of order $\frac{\widetilde{V}}{\sqrt{N(s,a)}}$)    ▷ Bonus scaling
    Run Value-Iteration with Projection (Algorithm 2).
    Set $\pi_t$ as the returned policy of VI.
End for
```

As in standard UCBVI, we define $N_h^t(s,a)$ to be the visitation count for the state-action pair (s,a) at iteration $t-1$ for horizon h: $N_h^t(s,a) := \sum_{i=0}^{t-1} \mathbf{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}$. Similarly to $N_h^t(s,a)$, we use $N_h^t(s,a,s') := \sum_{i=0}^{t-1} \mathbf{1}\left\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right\}$ as the visitation count of state-action pair (s,a) and the subsequent state s'. We then use $\widehat{\mathbb{P}}_t(s'|s,a) := \frac{N_h^t(s,a,s')}{N_h^t(s,a)}$,[3] $\forall h, s, a, s'$ to denote the empirical transition kernels at iteration t. UCBVI uses value iteration with the empirical transition function $\widehat{\mathbb{P}}_t$ and a reward function augmented with an exploration bonus, given by $r_h + b_h^t$. This is defined as a dynamic programming procedure that starts at H and then proceeds backward in time to $h = 0$, updating according to Algorithm 1 and 2. The key algorithmic changes between UCBVI and UCBVI-Shaped are highlighted in blue: (1) scaling bonus b_h^t by the shaping term $\widetilde{V}$ and (2) projecting the empirical value function $\widehat{V}_h^t(s)$ by an upper bound based on the sandwiched shaping, $\beta\widetilde{V}(s,a)$. As we discuss in Section 6, we also show how the sandwich factor β does need to be provided beforehand but instead can be inferred through a straightforward technique for online model selection. While the approximate order of the bouns term is $\frac{\widetilde{V}}{\sqrt{N(s,a)}}$, a more detailed description can be found in Section 5.

Algorithm 2 Value Iteration with Projection

```
Input $\left\{\widehat{P}_t, r + b_h^t\right\}_{h=0}^{H-1}$.
$\widehat{V}_H^t(s) = 0, \forall s, \widehat{Q}_h^t(s,a)$
While not converged
    $\widehat{Q}_h^t(s,a) = \min\left\{r_h(s,a) + b_h^t(s,a) + \widehat{\mathbb{P}}_t(\cdot|s,a) \cdot \widehat{V}_{h+1}^t, H\right\}$
    $\widehat{V}_h^t(s) = \min\left(\max_a \widehat{Q}_h^t(s,a), \beta\widetilde{V}(s)\right)$    ▷ Value projection
$\pi_h^t(s) = \arg\max_a \widehat{Q}_h^t(s,a), \forall h, s, a.$
```

5 Analyzing UCBVI-Shaped

In this section, we will derive our main result on the sample complexity of the UCBVI-Shaped algorithm, and along the way introduce pseudosuboptimal and path-pseudosuboptimal states as a tool for deriving bounds that depend only on the "effective" state space as determined by the provided

[3] The $\widehat{\mathbb{P}}_t$ here is dependent on the horizon h, but since we have assumed (Assumption 3) the states s are h-indexed, we will use $\widehat{\mathbb{P}}_t$ for notation simplicity.

reward shaping. We will first introduce some notation to use in our analysis. As described in the previous section, UCBVI-Shaped proceeds in rounds. At the beginning of round t, the learner has access to an empirical model $\widehat{\mathbb{P}}_t$ built from the data collected up to iteration $t-1$. The bonus terms we consider are built by taking the empirical second moment of $\widetilde{V}$. This is related to the definition of bonus$_2$ in Azar et al. [8]. Importantly, the empirical value functions $\widehat{V}_h^t$ are clipped above by $\beta\widetilde{V}_h$:

$$b_h^t(s,a) = \min\left(16\beta\sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}_{h+1}^2(s')|s,a]\ln\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}{N_h^t(s,a)}} + \frac{12\beta\widetilde{V}^{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}, 2\beta\widetilde{V}^{\max}\right), \quad (1)$$

where $\widetilde{V}_h^{\max} = \max_{s'}\widetilde{V}_h(s')$ and $\widetilde{V}^{\max} = \max_{s',h'}\widetilde{V}_{h'}(s')$.

Despite clipping and a modified bonus term, the $\widehat{Q}$ and $\widehat{V}$ values of UCBVI-Shaped are optimistic:

Lemma 5.1. *With probability at least $1-\delta$ we have*

$$\widehat{V}_0^t(s_0) \geq V_0^\star(s_0),\ \forall s_0 \in \mathcal{S}; \qquad \textit{and} \qquad \widehat{Q}_h^t(s,a) \geq Q_h^\star(s,a), \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}, \quad (2)$$

for all $t,h \in \mathbb{N}\times[H]$, where $\widehat{V}_h^t$ is computed via Algorithm 2.

The proof of Lemma 5.1 can be found in Appendix A.1. Optimism (Lemma 5.1) and the simulation Lemma [2] imply that:

$$V^\star(s_0) - V^{\pi_t}(s_0) \leq \widehat{V}_1^t(s_0) - V^{\pi_t}(s_0) = \mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^H b_h^t(s_h,a_h) + \left(\widehat{\mathbb{P}}_h^t(\cdot|s_h,a_h) - \mathbb{P}^\star(\cdot|s_h,a_h)\right)\cdot\widehat{V}_{h+1}^{\pi_t}\right]. \quad (3)$$

We now turn our attention to characterize the information contained in $\widetilde{V}$.

5.1 Pruning of the State Space

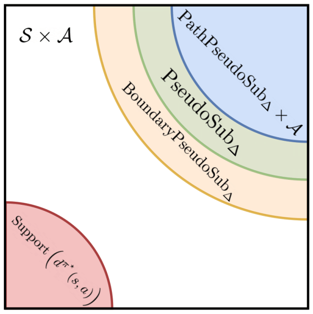

$\mathcal{S}\times\mathcal{A}$
PathPseudoSub$_\Delta$ + $\mathcal{A}$
PseudoSub$_\Delta$
BoundaryPseudoSub$_\Delta$
Support$\left(d^{\pi^\star}(s,a)\right)$

Figure 2: PathPseudoSub$_\Delta \times \mathcal{A}$ is split from Support$(d^{\pi^*}(s,a))$ by BoundaryPseudoSub$_\Delta$. UCBVI-Shaped can avoid exploring over PathPseudoSub$_\Delta \times \mathcal{A}$.

Consider the following "surrogate" Q functions, $\widetilde{Q}_h^u : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ induced by $\widetilde{V}$ via:

$$\widetilde{Q}_h^u(s,a) = \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[r(s,a) + \beta\tilde{V}_{h+1}(s')\right].$$

By Assumption 1, we can bound $Q^\star$ via $Q_h^\star(s,a) \leq \widetilde{Q}_h^u(s,a)$. The basis of our main results is the following observation. If an action a satisfies $\widetilde{Q}_h^u(s,a) < V_h^\star(s)$ for state s, then a is a suboptimal action for state s. The projection (Step 5 of Algorithm 2) ensures the 'empirical' $Q-$functions $\widehat{Q}_h^t(s,a)$ of Algorithm 1 will quickly converge to values upper bounded by $\widetilde{Q}_h^u(s,a)$. Since optimism guarantees that $\widehat{Q}_h^t(s,\pi_*(s)) \geq V^\star(s)$, and the policy executed by UCBVI-Shaped is greedy w.r.t $\widehat{Q}_h^t$, actions belonging to state-action pairs such that $\widetilde{Q}_h^u(s,a) < V_h^\star(s)$ will quickly stopped being played by the algorithm. Moreover, all states that are only accessible through state-action pairs of this kind will also stop being visited by the algorithm after only a few iterations. In the subsequent discussion we will call PseudoSub to the set of state action pairs that are quickly 'prunned out' by UCBVI-Shaped, the set of states only accessible through these state action pairs as PathPseudoSub and the set of state-action pairs in PseudoSub that are not in PathPseudoSub $\times$ $\mathcal{A}$ as BoundaryPseudoSub. Once all states in BoundaryPseudoSub have been visited enough, no states in PathPseudoSub will be visited again. Provided this happens sufficiently fast, we can show a regret bound that is independent on the size of PathPseudoSub. The subsequent discussion is aimed at formalizing this and culminates with our main result (Theorem 5.2).

Given a parameter $\Delta > 0$, we say that an action a is $\Delta-$pseudosuboptimal[4] for state s if $V_h^\star(s) \geq \Delta + \widetilde{Q}_h^u(s,a)$. We denote the set of $\Delta-$pseudosuboptimal state action pairs as:

$$\text{PseudoSub}_\Delta = \{(s,a)\in\mathcal{S}\times\mathcal{A} \text{ s.t. } V_h^\star(s) \geq \Delta + \widetilde{Q}_h^u(s,a)\}.$$

[4] We add the qualifier pseudo to our name to distinguish between suboptimality as captured by our surrogate Q values and true suboptimality between the true values of $Q_h^\star$.

The intuition we want to capture is that all states $\tilde{s}$ that can be accessed only through traversing a state action pair in PseudoSub_Δ, can be safely ignored because we can determine their suboptimality as soon as the state action pairs in PseudoSub_Δ that lead to $\tilde{s}$ are visited enough.

Now we define the set of Δ-path-pseudosuboptimal states as the states that can be reached only by traversing a $\Delta-$pseudosuboptimal state action pair:

$$\text{PathPseudoSub}_\Delta = \{s \in \mathcal{S} \text{ s.t. all feasible paths from initial states to } s \text{ intersect } \text{PseudoSub}_\Delta\}.$$

Notice that there may be state action pairs in $\text{PathPseudoSub}_\Delta \times \mathcal{A}$ that may not be in PseudoSub_Δ. In fact, there may exist states s in $\text{PathPseudoSub}_\Delta$ such that no (s,a) is in PseudoSub_Δ for all $a \in \mathcal{A}$. We will show that states in $\text{PathPseudoSub}_\Delta$ will not be explored by UCBVI-Shaped after a few iterations. Intuitively, this happens because the support of the state visitation distribution of the optimal policy does not contain any state in $\text{PathPseudoSub}_\Delta, \forall \Delta > 0$ (or equivalently, $\text{Support}(d^{\pi^\star}(s,a)) \cap \text{PathPseudoSub}_\Delta = \emptyset, \forall \Delta > 0$). Hence, once the UCBVI-shaped identifies some sub-optimal state action pairs, the algorithm will not visit these state-action pairs again.

For any state action pair (s,a), define its set of neighboring states $\text{Neighbor}(s,a)$ as the set of states with nonzero probability in $\mathbb{P}^\star(\cdot|s,a)$. By definition of $\text{PathPseudoSub}_\Delta$, for all $(s,a) \in (\mathcal{S} \times \mathcal{A}) \backslash ((\text{PathPseudoSub}_\Delta \times \mathcal{A}) \cup \text{PseudoSub}_\Delta)$, we have:

$$\text{Neighbor}(s,a) \subseteq \mathcal{S} \backslash \text{PathPseudoSub}_\Delta. \tag{4}$$

In other words, the neighborhood of any state action pair (s,a) whose state is not in $\text{PathPseudoSub}_\Delta$ and such that (s,a) is not $\Delta-$pseudo-suboptimal, are not in $\text{PathPseudoSub}_\Delta$. We also introduce the notion of "boundary pseudosuboptimal" state action pairs to capture the set of state action pairs that are $\Delta-$suboptimal but whose states are not in $\text{PathPseudoSub}_\Delta$.

$$\text{BoundaryPseudoSub}_\Delta = \{(s,a) \in \text{PseudoSub}_\Delta \text{ and } s \notin \text{PathPseudoSub}_\Delta\}.$$

Naturally, one of these states has to be traversed by any trajectory that contains any state in $\text{PathPseudoSub}_\Delta$.

Although we only use $\widetilde{Q}^u$ in the definition of PseudoSub_Δ, $\text{PathPseudoSub}_\Delta$, and $\text{BoundaryPseudoSub}_\Delta$, the size of these sets is modulated by the scale of β and the width of the intervals $\left[\widetilde{Q}^l(s,a), \widetilde{Q}^u(s,a)\right]$. We will show that (in the notation of Theorem 3.1) $\mathcal{S}_{\text{pruned}} \approx \mathcal{S} \backslash \text{PathPseudoSub}_\Delta$. With this notation, the formal version of our main results is stated as follows.

Theorem 5.2. *With probability at least $1 - 6\delta$, the regret of UCBVI-Shaped is upper bounded by*

$$\sum_{t=1}^{T} V^\star(s_0) - V^{\pi_t}(s_0) = \mathcal{O}\Bigg(\min_\Delta \Bigg(H\beta\widetilde{V}^{\max}\sqrt{|\mathcal{S}\backslash\text{PathPseudoSub}_\Delta||\mathcal{A}|T \ln\frac{\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|T}{\delta}} + \beta^2 \left(\widetilde{V}^{\max}\right)^2 H^{1/2} |\text{BoundaryPseudoSub}_\Delta|^{1/2} \ln\frac{\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|T}{\delta} \times \min\left(A(\Delta), B(\Delta)\right) \Bigg)\Bigg).$$

for all $T \in \mathbb{N}$. Where $A(\Delta) = \frac{|\mathcal{S}|^{1/2}|\mathcal{A}|^{1/2}}{\Delta}$ and $B(\Delta) = \frac{\beta\widetilde{V}^{\max}H^{1/2}|\text{BoundaryPseudoSub}_\Delta|^{1/2}}{\Delta^2}$.

Theorem 5.2 instantiates the desiderata of Theorem 3.1. Although this regret upper bound cannot be decomposed into a sum of two term as in Theorem 3.1. For any fixed Δ, the regret upper bound has two components, one where $\mathcal{S}_{\text{remain}}$ can be identified with $|\mathcal{S}\backslash\text{PathPseudoSub}_\Delta|$ and a second one that scales logarithmically in T. In the next section we flesh out the steps in the proof of this result. The full argument can be found in Appendix A. The bonus scaling also allows us to ameliorate the horizon dependence of the upper bound. Instead of obtaining a H^2 dependence as the bound of theorem 7.1 in [2], the first term depends on $H\beta\widetilde{V}^{\max}$. Notice that the state dependence in the second term may be only be of order $\log(|\mathcal{S}|)$ if there is a Δ such that $|\text{BoundaryPseudoSub}_\Delta| \ll |\mathcal{S}|$ (albeit at the cost of a quadratic dependence on $1/\Delta^2$). Moreover notice that $|\text{BoundaryPseudoSub}_\Delta|$ could be much smaller than $|\text{PathPseudoSub}_\Delta|$ (the set of states that are reachable only by visiting states in $|\text{PseudoSub}_\Delta|$) thus showing that in some cases we can guarantee that even in the low order terms, the regret of UCBVI-Shaped that has polynomial dependence on an effective state space size that may be orders of magnitude smaller than the original one. The full version of this bound, with all the low order terms we have omitted for the sake of readability can be found in Appendix B.1, Theorem B.11. Note that Theorem 5.2 is a strict generalization to the UCBVI regret upper bounds, as setting Δ to a value that is smaller than the minimum gap recovers the exact result (Theorem 7.6in [2]).

5.2 Proof Intuitions and Sketch for Theorem 5.2

Improved Horizon Dependence. An empirical Bernstein bound shows that adding a bonus scaling (up to low order terms) with $\mathcal{O}\left(\sqrt{\widehat{\mathrm{Var}}_{s'\sim\widehat{\mathbb{P}}^t_h(\cdot|s,a)}(V^\star_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}/N^t_h(s,a)}\right)$ is sufficient to ensure optimism. Since $V^\star$ is not known, this variance term can be substituted by the empirical second moment of $\beta\widetilde{V}_{h+1}$. Finally, the scaling of these terms can be upper bounded by $\beta\widetilde{V}^{\max}$. Without knowledge of $\widetilde{V}$, achieving this scaling would be challenging, since the only proxy for $V^\star_{h+1}$ available is $\widehat{V}^t_{h+1}$ or H both of which may vastly overestimate it.

State Pruning. The value function clipping mechanism ensures that $\mathbb{E}_{s'\sim\widehat{\mathbb{P}}^t_h(\cdot|s,a)}\left[\widehat{V}^t_{h+1}(s')\right] \leq \beta\mathbb{E}_{s'\sim\widehat{\mathbb{P}}^t_h(\cdot|s,a)}\left[\widetilde{V}_{h+1}(s')\right]$ and therefore the empirical gap between $\widehat{Q}^t_h(s,a)$ and $V^\star(s)$ is decreasing at a rate of at most $\mathcal{O}\left(\beta\widetilde{V}^{\max}/\sqrt{N^t_h(s,a)}\right)$. Since optimism ensures that $\widehat{Q}^t_h(s,\pi^\star(a)) \geq V^\star(s)$, once $\widehat{Q}(s,a) < V^\star(s)$, action a will not be chosen anymore. Thus, for any Δ, the number of times any state action pair in $\mathrm{PseudoSub}_\Delta$ may be visited by UCBVI-Shaped is upper bounded by $\beta^2\left(\widetilde{V}^{\max}\right)^2/\Delta^2$. Since any visit to a state action pair in $\mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta$ requires a visit to a state in $\mathrm{BoundaryPseudoSub}_\Delta$, which allows us to bound the total number of visits to a state (or a trajectory containing such a state) in $\mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta$ by $H|\mathrm{BoundaryPseudoSub}_\Delta|\beta^2\left(\widetilde{V}^{\max}\right)^2/\Delta^2$.

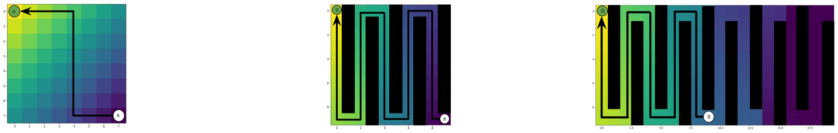
Finally, we can split (a version of) the regret decomposition in Eqn. 3 into two sums, one over state action pairs in $\mathcal{U}^{\mathrm{bad}} = \mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta$ (or trajectories intersecting $\mathcal{U}^{\mathrm{bad}}$) and a second one over state action pairs in $\mathcal{U}^{\mathrm{good}} = (\mathcal{S}\times\mathcal{A})\setminus(\mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta)$ (or trajectories without intersection with $\mathcal{U}^{\mathrm{bad}}$). We can then apply the upper bound on the visitation of state action pairs in $\mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta$ to derive a regret upper bound over these states scaling with $1/\Delta$. Using the bound on the number of trajectories that intersect $\mathrm{PathPseudoSub}_\Delta \times \mathcal{A} \cup \mathrm{PseudoSub}_\Delta$, the trajectory decomposition yields the term scaling with $1/\Delta^2$ (but having only logarithmic state dependence) in Theorem 5.2. Recall that Eqn. 4 implies that the transition operators over state action pairs in $\mathcal{U}^{\mathrm{good}}$ have support only over $\mathcal{S}\setminus\mathrm{PathPseudoSub}_\Delta$. Our proofs use this fact to prove a polynomial dependence on $|\mathcal{S}\setminus\mathrm{PathPseudoSub}_\Delta|$ and not $|\mathcal{S}|$ in Theorem B.11 (see Appendix B.1), the full version of Theorem 5.2. Detailed proofs are in Appendix A.

6 Practical Considerations: Online Model Selection

Now one may notice that UCBVI-Shaped requires knowledge of the scaling β is in order to actually perform the value projection. While this can be pre-provided by a user or set conservatively, in this section we discuss how this can be inferred by viewing this as an online model selection problem. In particular, given a set of N different values of β — $[\beta_1, \beta_2, \ldots, \beta_N]$, each of which parameterizes a different setting of the learning algorithm UCBVI-Shaped(β), an online model selection algorithm such as Stochastic CORRAL [1, 34] or RegretBalancing [33, 15] jointly infers the value of β and learns the appropriate value function online. In Appendix B.4 we show how these techniques can yield meaningful regret guarantees and we provide pseudocode.

7 Numerical Simulations

To show the practical relevance of our analysis on reward shaping we perform some numerical simulations on a family of maze environments with tabular state-action representations, as shown in Fig. 3. These environments have deterministic dynamics and reward. The reward is 0 at all states except a goal sink state, which has reward 1. These simulations are aimed at studying the following questions: **(1)** Does reward shaping improve sample complexity in these types of maze environments over uninformed UCBVI? **(2)** What is the relative importance of the bonus reweighting and the $\widetilde{V}$ projection? **(3)** How does the "suboptimality" of $\widetilde{V}$ impact the resulting sample complexity? **(4)** Does introducing decayed shaping actually allow for policy convergence? In these experiments, $\widetilde{V}$ is constructed by scaling the optimal value function $V^\star$ by per-state scaling factors sampled independently within the range β.

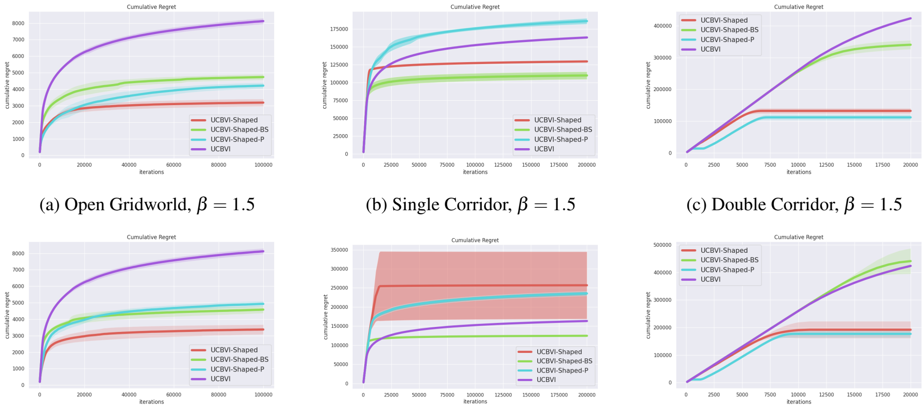


(a) Value function of Gridworld (b) Value function of Single Corridor (c) Value function of Double Corridor

Figure 3: Environments used for numerical simulations. **(Left)** Open Gridworld **(Middle)** Single corridor, agent starts bottom right has to reach a goal in top left **(Right)** Double corridor, agent starts in the middle and has to reach a goal on the left, with many irrelevant states on the right hand side

7.1 Does reward shaping help direct exploration over optimism under uncertainty?

We conducted numerical simulations on tabular environments to understand how reward shaping via UCBVI-Shaped provides benefits over standard UCBVI. Fig. 4 shows cumulative regret accumulated with different variants of UCBVI-Shaped (with both projection and bonus scaling), UCBVI-Shaped-P (with only projection), UCBVI-Shaped-BS (with only bonus scaling), UCBVI (standard UCBVI without shaping, as described in [8]). This is benchmarked across the three environments described in Fig. 3, with various levels of imperfect shaping applied by varying $\beta = \{1.5, 1.9\}$. As seen from Fig. 4, across all environments UCBVI-shaped with projection and bonus scaling performs most favorably, followed typically by UCBVI-Shaped-P, followed by UCBVI-Shaped-BS and UCBVI, suggesting that reward shaping can significantly help with learning efficiency.

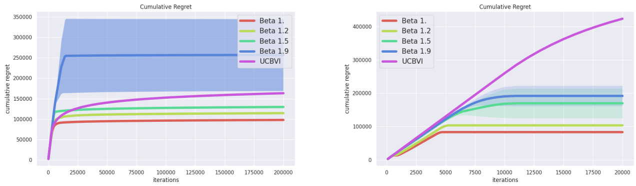


(a) Open Gridworld, $\beta = 1.5$ (b) Single Corridor, $\beta = 1.5$ (c) Double Corridor, $\beta = 1.5$

(d) Open Gridworld, $\beta = 1.9$ (e) Single Corridor, $\beta = 1.9$ (f) Double Corridor, $\beta = 1.9$

Figure 4: Cumulative regret for learning in various environments with varying amounts of shaping, as compared with UCBVI, and ablations UCBVI-Shaped-BS (no projection) and UCBVI-Shaped-P (no bonus scaling).

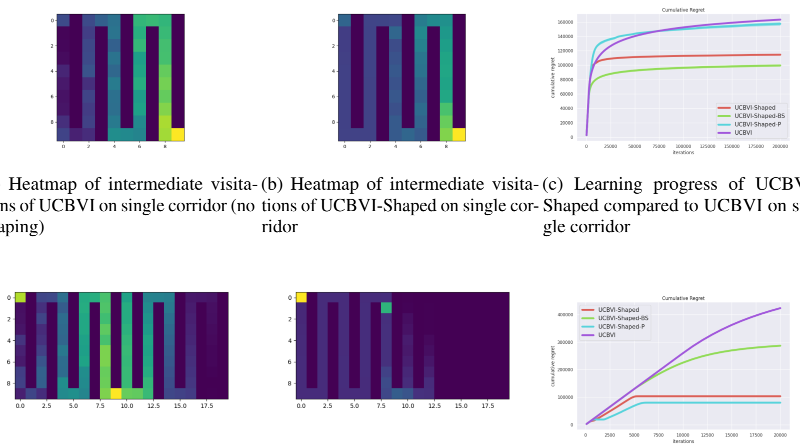
7.2 How does the effectiveness of reward shaping vary across environments?



(a) Single Corridor (b) Double Corridor

Figure 5: Effect of suboptimality of reward shaping on the performance of UCBVI-Shaped. While $\beta = 1.2, 1.5$ don't make much of a difference, very large β leads to performance degradation

We next conducted some numerical simulations across environments to understand how the nature of the environment itself affects the sample complexity of learning with shaping via UCBVI-Shaped. As shown in Fig. 6, we see that UCBVI shaped can be very effective in environments with many irrelevant sub-optimal paths like the double corridor environment in Fig. 3, but is relatively less effective in environments where all exploration is directed the same way such as the single corridor. Even incorrect but optimistic shaping will provide guidance towards the goal, making UCBVI relatively less dominant in the single corridor environment as compared to the double corridor. This suggests that in environments where



(a) Heatmap of intermediate visitations of UCBVI on single corridor (no shaping)

(b) Heatmap of intermediate visitations of UCBVI-Shaped on single corridor

(c) Learning progress of UCBVI-Shaped compared to UCBVI on single corridor

(d) Heatmap of intermediate visitations of UCBVI on double corridor (no shaping)

(e) Heatmap of intermediate visitations of UCBVI-Shaped on double corridor

(f) Learning progress of UCBVI-Shaped compared to UCBVI on double corridor

Figure 6: Visualization of how different environments are affected by reward shaping differently. (Left) intermediate visitations (Right) learning progress of UCBVI vs UCBVI-Shaped. The single corridor environment on the top sees much smaller gains for UCBVI-Shaped compared to double corridor environment.

ruling out an entire part of the exploration space is easy from the shaping, we can expect to see larger benefits.

7.3 How does the suboptimality of reward shaping affect learning?

We next compared how different levels of suboptimality of the β sandwich term in the reward shaping affect cumulative regret across environments. As shown in Fig. 4 (a) and (d), we see that for environments with open paths (like the open gridworld), the shaping degradation has minimal negative effect until it gets very suboptimal. On the other hand, for corridor and double corridor (Fig. 5 (b)), where there are only a few paths to the goal, suboptimal reward shaping along those paths significantly hamper progress.

7.4 Is online UCBVI-Shaped able to infer β online without prior knowledge?

As described in Section 6, UCBVI-shaped can be freed of the assumption of β being known by performing online model selection of β and learning values jointly. In particular, we use the Stochastic CORRAL algorithm [34], a variant of the method introduced in [1] to perform online model selection, with the episodic return being the requisite criterion for updating the model selection distribution. As we see in Fig 7, this scheme is able to show comparable results to when the actual β is known beforehand, only degrading as the value of β increases. This suggests that online UCBVI-shaped can be practical in regimes with moderate levels of value corruption.

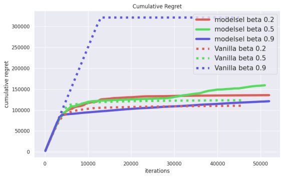


Figure 7: Understanding performance of online model selection in UCBVI-Shaped

8 Discussion

In this work, we take a step towards formally analyzing the benefits of reward shaping, proving that effective reward shaping can lead to more efficient learning than uninformed exploration strategies. In our analysis, we study an algorithm that incorporates reward shaping into a modified version of UCBVI, using it to modify bonuses and value function projection. Our analysis shows that incorporating shaped rewards allows for pruning significant parts of the state space and sharpening of optimism in a task directed way. This reduces the dependence of the regret bound on the state space size and on the horizon, depending on the quality of the shaping term and parameters of the MDP. This shows how reward shaping can direct exploration and provide significant sample complexity benefits while retaining asymptotic performance. We hope that this work is a step towards moving sample complexity analysis away from being reward agnostic to actually considering reward shaping more formally in analysis.

References

[1] A. Agarwal, H. Luo, B. Neyshabur, and R. E. Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

[2] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, 2019.

[3] A. Agarwal, M. Henaff, S. Kakade, and W. Sun. Pc-pg: Policy cover directed exploration for provable policy gradient learning. *Advances in Neural Information Processing Systems*, 33: 13399–13412, 2020.

[4] S. Agrawal and R. Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

[5] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[6] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

[7] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[8] M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

[9] Y. Bai, T. Xie, N. Jiang, and Y.-X. Wang. Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

[11] 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.

[12] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[13] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[14] C. Cheng, A. Kolobov, and A. Swaminathan. Heuristic-guided reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 13550–13563, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/70d31b87bd021441e5e6bf23eb84a306-Abstract.html.

[15] A. Cutkosky, C. Dann, A. Das, C. Gentile, A. Pacchiano, and M. Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pages 2276–2285. PMLR, 2021.

[16] C. Dann, T. V. Marinov, M. Mohri, and J. Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 34:1–12, 2021.

[17] Y. Efroni, N. Merlis, M. Ghavamzadeh, and S. Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] A. Faust, K. Oslund, O. Ramirez, A. G. Francis, L. Tapia, M. Fiser, and J. Davidson. PRM-RL: long-range robotic navigation tasks by combining reinforcement learning and sampling-based planning. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018, Brisbane, Australia, May 21-25, 2018*, pages 5113–5120. IEEE, 2018. doi: 10.1109/ICRA.2018.8461096. URL https://doi.org/10.1109/ICRA.2018.8461096.

[19] D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[20] R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

[21] N. Golowich and A. Moitra. Can q-learning be improved with advice? In *Conference on Learning Theory*, pages 4548–4619. PMLR, 2022.

[22] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[23] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

[24] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

[25] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[26] S. M. Kakade et al. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

[27] G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[28] K. Li, A. Gupta, A. Reddy, V. H. Pong, A. Zhou, J. Yu, and S. Levine. Mural: Meta-learning uncertainty-aware rewards for outcome-driven reinforcement learning. In *International Conference on Machine Learning*, pages 6346–6356. PMLR, 2021.

[29] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[30] P. Ménard, O. D. Domingues, X. Shang, and M. Valko. Ucb momentum q-learning: Correcting the bias without forgetting. In *International Conference on Machine Learning*, pages 7609–7618. PMLR, 2021.

[31] A. Y. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[32] A. Pacchiano, P. Ball, J. Parker-Holder, K. Choromanski, and S. Roberts. On optimism in model-based reinforcement learning. *arXiv preprint arXiv:2006.11911*, 2020.

[33] A. Pacchiano, C. Dann, C. Gentile, and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

[34] A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems*, 33:10328–10337, 2020.

[35] A. Pacchiano, P. Ball, J. Parker-Holder, K. Choromanski, and S. Roberts. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR, 2021.

[36] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

[37] R. Raileanu and T. Rocktäschel. Ride: Rewarding impact-driven exploration for procedurally-generated environments. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkg-TJBFPB.

[38] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[39] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550 (7676):354–359, 2017.

[40] B. C. Stadie, S. Levine, and P. Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

[41] H. Tang, R. Houthooft, D. Foote, A. Stooke, O. X. Chen, Y. Duan, J. Schulman, F. DeTurck, and P. Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pages 2753–2762, 2017.

[42] H. Van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang. Hybrid reward architecture for reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/1264a061d82a2edae1574b07249800d6-Paper.pdf.

[43] A. S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, and K. Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3540–3549. JMLR. org, 2017.

[44] 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*, 575(7782):350–354, 2019.

[45] Y. Wang, R. Wang, S. S. Du, and A. Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

[46] K. Yang, L. Yang, and S. Du. Q-learning with logarithmic regret. In *International Conference on Artificial Intelligence and Statistics*, pages 1576–1584. PMLR, 2021.

[47] L. Yang and M. Wang. Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*, pages 6995–7004. PMLR, 2019.

[48] L. Yang and M. Wang. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pages 10746–10756. PMLR, 2020.

[49] Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces. *arXiv preprint arXiv:2011.04622*, 2020.

[50] D. Ye, G. Chen, W. Zhang, S. Chen, B. Yuan, B. Liu, J. Chen, Z. Liu, F. Qiu, H. Yu, Y. Yin, B. Shi, L. Wang, T. Shi, Q. Fu, W. Yang, L. Huang, and W. Liu. Towards playing full moba games with deep reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 621–632. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/06d5ae105ea1bea4d800bc96491876e9-Paper.pdf.

[51] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In L. P. Kaelbling, D. Kragic, and K. Sugiura, editors, *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, pages 1094–1100. PMLR, 2019. URL `http://proceedings.mlr.press/v100/yu20a.html`.

[52] 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*, pages 7304–7312. PMLR, 2019.

[53] A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

[54] Y. Zhai, C. Baek, Z. Zhou, J. Jiao, and Y. Ma. Computational benefits of intermediate rewards for goal-reaching policy learning. *Journal of Artificial Intelligence Research*, 73:847–896, 2022.

[55] Z. Zhang, Y. Zhou, and X. Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.

[56] D. Zhou, J. He, and Q. Gu. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pages 12793–12802. PMLR, 2021.

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]
   (b) Did you describe the limitations of your work? [Yes] See Section 8
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 8
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] This work does not actually use human subjects, and is largely done in simulation. But we have included a discussion in Section 8
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] Math is used as a theory/formalism, but we don't make any provable claims about it.
   (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] See Appendix A for link to URL and run instructions in the README in the github repo.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix A.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All plots were created with 3 random seeds with std error bars.
   (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] See Appendix A
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] Envs we used are cited in section 7
   (b) Did you mention the license of the assets? [Yes] This is in Appendix B

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We published the code and included all environments and assets as a part of this

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used three open source domains and collected our own data on these domains.

(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? [No] We did not include full text since we didn't use an exact script, but we summarized the instructions and included images of the environments used.

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] Only human involvement was data collection with our system.

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

A Full Proofs

Let $f_h : \mathcal{S} \to \mathbb{R}$ be an arbitrary family of horizon indexed functions satisfying $\|f_h\|_\infty \leq B$ for some $B > 0$. We assume f_h is not a random variable as a function of $\mathcal{S}$. The following upper bounds hold,

Lemma A.1. *Fix $\delta \in (0,1)$, $\forall t \in \mathbb{N}$, $s \in \mathcal{S}$, $a \in \mathcal{A}$ and $h \in [H]$, with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \leq 16\sqrt{\frac{\mathrm{Var}(f_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)} + \frac{12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}}$$
$$\leq 16B\sqrt{\frac{\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)} + \frac{12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}}$$

for all $t \in \mathbb{N}$. Similarly with probability at least $1-\delta$, we have

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \leq 16\sqrt{\frac{\widehat{\mathrm{Var}}(f_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)} + \frac{12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}}$$

for all $t \in \mathbb{N}$. Where $\widehat{\mathrm{Var}}(f_{h+1}(s')|s,a) = \frac{1}{N_h^k(s,a)(N_h^k(s,a)-1)}\sum_{1\leq i<j<N_h^k(s,a)}(f_{h+1}(s_{h+1}^i) - f_{h+1}(s_{h+1}^j))^2$.

Proof. Consider a fixed tuple $s,a,t,h \in \mathcal{S}\times\mathcal{A}\times\mathbb{N}\times[H]$. By the definition of $\widehat{\mathbb{P}}_t$, we have

$$\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} = \frac{1}{N_h^t(s,a)}\sum_{i=1}^{t-1}\mathbf{1}\left\{(s_{h'}^i, a_{h'}^i) = (s,a)\right\} f_{h+1}(s_{h'+1}^i).$$

Now denote $\mathcal{H}_{h,i}$ as the entire history from $t=0$ to iteration $t=i$ where in iteration i, the history $\mathcal{H}_{h,i}$ includes all interactions from step 0 up to and including time step h. Next, $\forall i = 0,1,\ldots,t-1$, define the random variables:

$$X_{i,h'}(s,a) = \mathbf{1}\left\{(s_{h'}^i,a_{h'}^i)=(s,a)\right\}f_{h+1}(s_{h'+1}^i) - \mathbb{E}\left[\mathbf{1}\left\{(s_{h'}^i,a_{h'}^i)=(s,a)\right\}f_{h+1}(s_{h'+1}^i)|\mathcal{H}_{h,i}\right].$$

Notice that $\left|X_{i,h'}\right| \leq B$, if $\mathbf{1}\left\{(s_{h'}^i,a_{h'}^i)=(s,a)\right\}=1$, else $\left|X_{i,h'}\right|=0$ so that

$$\mathrm{Var}(X_{t,h}(s,a)|H_{h-1,t}) \leq \begin{cases} 0 & \text{if } \left|X_{i,h'}\right| = 0 \\ B^2 & \text{o.w.} \end{cases}$$

An anytime Bernstein bound (see Lemma C.2) implies,

$$\left|\sum_{i=1}^{t-1}X_{i,h'}(s,a)\right| \leq 16\sqrt{N_h^t(s,a)\mathrm{Var}(f_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}} + 12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}$$
$$\leq 16B\sqrt{N_h^t(s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}} + 12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}.$$

With probability at least $1-\delta$ for all $t \in \mathbb{N}$. Thus, we have

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \leq 16B\sqrt{\frac{\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)} + \frac{12B\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}}$$

with probability at least $1-\delta$ for all $t\in\mathbb{N}$. The second inequality can be proven following the exact same proof process as described above, but instead making use of the empirical Bernstein bound of C.3 instead.

□

The following bound relates the error of $\widehat{\mathbb{P}}_t$ and $\mathbb{P}^\star$.

Corollary A.2 (State-action wise model error under $V^\star$ (Lemma 7.3 of [2]))**.** *Fix $\delta \in (0,1)$, $\forall t \in [1,2,\dots T], s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, consider $\forall V_h^\star : \mathcal{S} \to [0,H]$. With probability at least $1-\delta$, we have*

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top V^\star_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top V^\star_{h+1}\right| \leq \min\left(16\sqrt{\frac{\widehat{\mathrm{Var}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}(V^\star_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} + \frac{12V^\star_{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}, 2V^\star_{\max}\right),$$

for all $t \in \mathbb{N}$. Where $\widehat{\mathrm{Var}}(V^\star_{h+1}(s')|s,a) = \frac{1}{N_h^k(s,a)(N_h^k(s,a)-1)}\sum_{1\leq i<j<N_h^k(s,a)}(V^\star_{h+1}(s^i_{h+1}) - V^\star_{h+1}(s^j_{h+1}))^2$ and $V^\star_{\max} = \max_s V^\star(s)$.

Proof. A direct application of Lemma A.1 using $B = V^\star_{\max}$ yields the desired result. □

A direct consequence of Lemma A.2 and Assumption 1 is that,

Corollary A.3. *The $V^\star$ "variance bonus" is upper bounded by the empirical $\widetilde{V}$ "second moment bonus":*

$$16\sqrt{\frac{\widehat{\mathrm{Var}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}(V^\star_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} \leq 16\beta\cdot\sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}^2_{h+1}(s')|s,a]\ln\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}{N_h^t(s,a)}}$$

and

$$\frac{12V^\star_{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta} \leq \frac{12\beta\widetilde{V}^{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}.$$

Proof. For any random variable $X \sim \mathbb{P}$,

$$\mathrm{Var}(X) \leq \mathbb{E}_{X\sim\mathbb{P}}\left[X^2\right].$$

Therefore, for all s,a and any empirical distribution $\widehat{\mathbb{P}}_t(\cdot|s,a)$,

$$\widehat{\mathrm{Var}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}(V^\star_{h+1}(s')|s,a) \leq \widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}\left[\left(V^\star_{h+1}(s')\right)^2|s,a\right].$$

Finally, by Assumption 1,

$$\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}\left[\left(V^\star_{h+1}(s')\right)^2|s,a\right] \leq \beta^2\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}\left[\widetilde{V}^2_{h+1}(s')|s,a\right].$$

The result follows. □

Corollary A.3 and Lemma A.1 implies that with probability $1-\delta$ for all $t \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}$ and $h \in [1,\cdots,H]$ the bonuses b_h^t from Equation 1 satisfy,

$$\begin{aligned}\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right)\cdot V^\star_{h+1}\right| &\overset{(i)}{\leq} \min\left(16\sqrt{\frac{\widehat{\mathrm{Var}}(V^\star_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} + \frac{12V^\star_{\max}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}, 2V^\star_{\max}\right)\\ &\overset{(ii)}{\leq} b_h^t(s,a).\end{aligned} \tag{5}$$

Where inequality (i) follows by Corollary A.2 and (ii) by Corollary A.3.

Define as $\widehat{V}_h^t$ to the stage h value function of π_t as computed in the learned model and using bonus augmented rewards (with bonuses defined as in 1) with the data collected up to round $t-1$. We start by showing $\widehat{V}_h^t$ is always larger than $V_h^\star(s)$.

We show the following supporting Lemma, a generalization of Lemma A.1 where the functions $f_h : \mathcal{S} \to \mathbb{R}$ satisfying $\|f_h\|_\infty$ are allowed to be random. This lemma is a slight refinement from its corresponding result in the literature, where the dependence on the size of the support of $|\mathbb{P}^\star(\cdot|s,a)|$ is upper bounded by $|\mathcal{S}$. Having a more refined control on the size of the support of $\mathbb{P}^\star(\cdot|s,a)$ is what allows our final bounds to only depend on the size of the effective state space.

Lemma A.4. *Fix $\delta \in (0,1)$, $\forall t \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right|$$
$$\le 16\beta \cdot \sqrt{\frac{|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}^2_{h+1}(s')|s,a]\ln\frac{8B|\mathcal{S}||\mathcal{A}|t^3}{\delta}}{N_h^t(s,a)}}$$
$$+ \quad \frac{12B|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|}{N_h^t(s,a)}\ln\frac{8B|\mathcal{S}||\mathcal{A}|t^3}{\delta} + \frac{1}{t^2}$$

simultaneously for all $\{f_h : \mathcal{S} \to \mathbb{R}\}_{h=0}^{H-1}$ such that $f_h(s) \in \left[\widetilde{V}_h, \beta\widetilde{V}_h\right]$ and $\|f_h\|_\infty \le B$. Where $\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))$ *corresponds to the size of the support of distribution* $\mathbb{P}^\star(\cdot|s,a)$.

Proof. The same proof template as in Lemma A.1 plus a union bound over a covering of the space of f functions yields the desired result. Note that for all $s,a \in \mathcal{S}\times\mathcal{A}$,

$$\widehat{\mathrm{Var}}(f_{h+1}(s')|s,a) \le \widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[f^2_{h+1}(s')|s,a] \le \beta^2\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}^2_{h+1}(s')|s,a]$$

Therefore Lemma A.1 implies that for any fixed f_{h+1}, with probability at least $1-\delta'$

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \tag{6}$$
$$\le 16\beta \cdot \sqrt{\frac{\widehat{\mathrm{Var}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}(f_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta'}}{N_h^t(s,a)} + \frac{12B}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta'}}$$
$$\le 16\beta \cdot \sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}^2_{h+1}(s')|s,a]\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta'}}{N_h^t(s,a)} + \frac{12B}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta'}} \tag{7}$$

for all (fixed) $f \in \{f_h : \mathcal{S} \to \mathbb{R}\}_{h=0}^{H-1}$ and for all $s,a,h \in \mathcal{S}\times\mathcal{A}\times H$ simultaneously. We now apply a standard $\varepsilon-$net covering argument to the inequality we have proven above. Notice that regardless of the number of samples, $\mathrm{support}(\widehat{\mathbb{P}}_t(\cdot|s,a)) \subseteq \mathrm{support}(\mathbb{P}^\star(\cdot|s,a))$. Thus, the only "entries" that matter in f_{h+1} are those corresponding to states in $\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))$.

Let's consider an $\varepsilon-$net of f_{h+1} restricted to $\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))$. Since for any $s \in \mathcal{S}$, we assume $f_{h+1} \in [\widetilde{V}_{h+1}, \beta\widetilde{V}_{h+1}]$, there exists an epsilon net $\mathcal{N}_\varepsilon$ under the infinity norm satisfying $|\mathcal{N}_\varepsilon| \le (\frac{2B}{\varepsilon})^{|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|}$. For any f_{h+1} we denote by f^ε_{h+1} its closest element in $\mathcal{N}_\varepsilon$ (in the infinity norm over $\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))$). The following holds,

$$\left|\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| - \left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f^\varepsilon_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f^\varepsilon_{h+1}\right|\right| \le$$
$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1}\right| + \left|\mathbb{P}^\star(\cdot|s,a)^\top f^\varepsilon_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f^\varepsilon_{h+1}\right| \le 2\varepsilon.$$

And therefore, setting $\delta' = \frac{\delta}{(\frac{2B}{\varepsilon})^{|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|}} \le \frac{\delta}{|\mathcal{N}_\varepsilon|}$, a union bound over all elements of $\mathcal{N}_\varepsilon$ implies that with probability at least $1-\delta$, we have

$$\begin{aligned}&\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \\ &\le 16\beta\cdot\sqrt{\frac{|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\tilde{V}^2_{h+1}(s')|s,a]\ln\frac{4B|\mathcal{S}||\mathcal{A}|t}{\varepsilon\delta}}{N_h^t(s,a)}} \\ &\quad+\frac{12B|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|}{N_h^t(s,a)}\ln\frac{4B|\mathcal{S}||\mathcal{A}|t}{\varepsilon\delta}+2\varepsilon\end{aligned}$$

for all f_{h+1} simultaneously. Setting $\varepsilon=\frac{1}{2t^2}$ we get that with probability at least $1-\delta$,

$$\begin{aligned}&\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top f_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top f_{h+1}\right| \\ &\le 16\beta\cdot\sqrt{\frac{|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\tilde{V}^2_{h+1}(s')|s,a]\ln\frac{8B|\mathcal{S}||\mathcal{A}|t^3}{\delta}}{N_h^t(s,a)}} \\ &\quad+\frac{12B|\mathrm{support}(\mathbb{P}^\star(\cdot|s,a))|}{N_h^t(s,a)}\ln\frac{8B|\mathcal{S}||\mathcal{A}|t^3}{\delta}+\frac{1}{t^2}\end{aligned}$$

for all f_{h+1} simultaneously and for all $t\in\mathbb{N}$. This completes the result.

□

A.1 Proof of Lemma 5.1

In this section we show that optimism holds for all state-action pairs. We restate Lemma 5.1 for the reader's convenience.

Lemma 5.1. *With probability at least* $1-\delta$ *we have*

$$\widehat{V}_0^t(s_0)\ge V_0^\star(s_0),\ \forall s_0\in\mathcal{S};\qquad \textit{and}\qquad \widehat{Q}_h^t(s,a)\ge Q_h^\star(s,a),\quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}, \tag{2}$$

for all $t,h\in\mathbb{N}\times[H]$, *where* $\widehat{V}_h^t$ *is computed via Algorithm 2.*

Proof. We prove via induction. At the additional time step H we have $\widehat{V}_H^t(s)=V_H^\star(s)=0$ for all s.

Starting at $h+1$, and assuming we have $\widehat{V}_{h+1}^t(s)\ge V_{h+1}^\star(s)$ for all s, we move to h below.

Consider any $s,a\in\mathcal{S}\times\mathcal{A}$. First if $\widehat{Q}_h^t(s,a)=H$ then we have $\widehat{Q}_h^t(s,a)\ge Q_h^\star(s,a)$. The following inequalities hold,

$$\begin{aligned}\widehat{Q}_h^t(s,a)-Q_h^\star(s,a)&=b_h^t(s,a)+\widehat{\mathbb{P}}_t(\cdot|s,a)\cdot\widehat{V}_{h+1}^t-\mathbb{P}^\star(\cdot|s,a)\cdot V_{h+1}^\star\\&\overset{(i)}{\ge} b_h^t(s,a)+\widehat{\mathbb{P}}_t(\cdot|s,a)\cdot V_{h+1}^\star-\mathbb{P}^\star(\cdot|s,a)\cdot V_{h+1}^\star\\&=b_h^t(s,a)+\left(\widehat{\mathbb{P}}_t(\cdot|s,a)-\mathbb{P}^\star(\cdot|s,a)\right)\cdot V_{h+1}^\star\\&\overset{(ii)}{\ge} b_h^t(s,a)-16\sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}^2_{h+1}(s')|s,a]\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}}-\frac{12\widetilde{V}^{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}\\&\overset{(iii)}{\ge}0\end{aligned}$$

Where (i) holds because of the inductive optimism assumption. Inequality (ii) holds because by Corollary A.2 we have,

$$\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right)\cdot V^\star_{h+1} \le 16\sqrt{\frac{\widehat{\mathrm{Var}}(V^\star_{h+1}(s')|s,a)\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} + \frac{12V^\star_{\max}\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}$$

and (iii) by Corollary A.3 applied to the definition of b_h^t.

□

B Supporting Results for Section 5.1

A consequence of Lemma A.1, and Corollary A.3,

Corollary B.1. *Fix $\delta \in (0,1)$, $\forall t \in \mathbb{N}, s \in \mathcal{S}, a \in \mathcal{A}, h \in [H]$, consider $\forall\, \widetilde{V}_h : \mathcal{S} \mapsto [0,H]$ with probability at least $1-\delta$, we have*

$$\left|\widehat{\mathbb{P}}_t(\cdot|s,a)^\top \widetilde{V}_{h+1} - \mathbb{P}^\star(\cdot|s,a)^\top \widetilde{V}_{h+1}\right| \le 16\sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}\left[\widetilde{V}^2_{h+1}(s')|s,a\right]\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} + \frac{12\widetilde{V}^{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}.$$

We will now show the empirical $\widetilde{Q}$ values can be approximately upper bounded by the $\widetilde{Q}^u$ values. Let's define the "tilde bonus" as $\tilde{b}_h^t(s,a) = 16\beta\sqrt{\frac{\widehat{\mathbb{E}}_{s'\sim\widehat{\mathbb{P}}_t(\cdot|s,a)}\left[\widetilde{V}^2_{h+1}(s')|s,a\right]\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_h^t(s,a)}} + \frac{12\beta\widetilde{V}^{\max}}{N_h^t(s,a)}\ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}$.

Corollary B.2. *With probability at least $1-2\delta$ the empirical Q function of state action pair (s,a) satisfies,*

$$\widehat{Q}_h^t(s,a) \le r(s,a) + b_h^t(s,a) + \tilde{b}_h^t(s,a) + \beta\mathbb{P}^\star(\cdot|s,a)^\top\widetilde{V}_{h+1} := \widetilde{Q}^u(s,a) + b_h^t(s,a) + \tilde{b}_h^t(s,a).$$

for all $s,a \in \mathcal{S}\times\mathcal{A}$ and for all $h \in [0,\cdots,H-1]$ and all $t \in \mathbb{N}$.

Proof. By definition $\widehat{Q}_h^t(s,a) = r(s,a) + b_h^t(s,a) + \widehat{\mathbb{P}}^t(\cdot|s,a)\widehat{V}_{h+1}^t$, since $\widehat{V}_{h+1}^t$ is clipped above by $\beta\widetilde{V}_{h+1}$,

$$\begin{aligned}
\widehat{Q}_h^t(s,a) &= r(s,a) + b_h^t(s,a) + \widehat{\mathbb{P}}^t(\cdot|s,a)\widehat{V}_{h+1}^t \\
&\le r(s,a) + b_h^t(s,a) + \beta\widehat{\mathbb{P}}^t(\cdot|s,a)\widetilde{V}_{h+1} \\
&\le r(s,a) + b_h^t(s,a) + \tilde{b}_h^t(s,a) + \beta\mathbb{P}^\star(\cdot|s,a)^\top\widetilde{V}_{h+1} \\
&= \widetilde{Q}^u(s,a) + b_h^t(s,a) + \tilde{b}_h^t(s,a).
\end{aligned}$$

The last inequality is a consequence of Corollary B.1. The result follows.

□

Corollary B.2 implies that once $b_h^t(s,a) + \tilde{b}_h^t(s,a) < V^\star(s) - \widetilde{Q}^u(s,a)$ optimism guarantees that from that point on, action a will never again be taken at state s. Indeed, since optimism (see Lemma 5.1) $\widehat{Q}_h^t(s,\pi_\star(s))$ is at least $V^\star(s)$, action a will be dominated by action $\pi_\star(s)$ from that point onward. We make this intuition precise in the following Lemma by upper bounding the number of times action a is taken at state s when $(s,a) \in \mathrm{PseudoSub}_\Delta$.

Lemma B.3. *With probability at least $1-2\delta$ and for all $\Delta \in [0,1]$ the state action pairs $(s,a) \in \mathrm{PseudoSub}_\Delta$ satisfy the bound,*

$$N_{h(s)}^t(s,a) \le \frac{8192\beta^2 \times \left(\widetilde{V}^{\max}\right)^2 \cdot \ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{\Delta^2}$$

For all $t \in \mathbb{N}$. Where $h(s)$ corresponds to horizon index of the state partitions that contains[5] *state s.*

[5] Recall that by Assumption 3 the states are assumed to be $h-$indexed and therefore the state space can be written as $\mathcal{S} = \cup_{h\in[H]}\mathcal{S}_h$. This also implies that $N_h^t(s) = 0$ for all $h \neq h(s)$.

Proof. We start by assuming the events of Lemma 5.1 (optimism) and Corollary B.2 ($\widehat{Q}_h^t(s,a) \leq \widetilde{Q}^u(s,a) + b_h^t(s,a) + \tilde{b}_h^t(s,a)$) hold. We do not need any $\Delta-$dependent high probability event to hold for our results to be valid.

Once $b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a) < \widetilde{V}(s) - \widetilde{Q}^u(s,a)$, action a will never be taken again at state s. The following bound holds for $b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a)$,

$$\begin{aligned} b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a) &\leq 32\beta \sqrt{\frac{\widehat{\mathbb{E}}_{s' \sim \widehat{\mathbb{P}}_t(\cdot|s,a)}[\widetilde{V}_{h(s)+1}^2(s')|s,a] \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_{h(s)}^t(s,a)}} + \frac{24\beta \max_{s'} \widetilde{V}_{h(s)+1}(s')}{N_{h(s)}^t(s,a)} \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta} \\ &\leq 32\beta \max_s \widetilde{V}_{h(s)+1}(s) \sqrt{\frac{\ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{N_{h(s)}^t(s,a)}} + \frac{24\beta \max_{s'} \widetilde{V}_{h(s)+1}(s')}{N_{h(s)}^t(s,a)} \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}. \end{aligned}$$

Assume $(s,a) \in \mathrm{PseudoSub}_\Delta$. When

$$N_{h(s)}^t(s,a) > \frac{8192\beta^2 \times \max_{s',h'} \widetilde{V}_{h'}^2(s') \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{\Delta^2}$$

we get

$$b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a) < \Delta \leq V^\star(s) - \widetilde{Q}^u(s,a). \tag{8}$$

The result follows because as soon as $b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a) < \Delta$ holds, the optimism guarantees that from that point on action a will never again be taken at state s. Indeed, let $\pi^\star(s)$ be the optimal action at state s,

$$\widehat{Q}_{h(s)}^t(s,\pi^\star(s)) \geq Q^\star(s,\pi^\star(s)) = V^\star(s) \geq \widetilde{Q}^u(s,a) + \Delta$$

Where the last inequality holds because by definition $(s,a) \in \mathrm{PseudoSub}_\Delta$.

Plugging in the bound of Corollary B.2, and invoking inequality 8,

$$\widehat{Q}_{h(s)}^t(s,\pi^\star(s)) \geq \widetilde{Q}^u(s,a) + \Delta > \widetilde{Q}^u(s,a) + b_{h(s)}^t(s,a) + \tilde{b}_{h(s)}^t(s,a) \geq \widehat{Q}_{h(s)}^t(s,a).$$

implying action a will not be selected by the greedy policy induced by the empirical Q function $\widehat{Q}_{h(s)}^t$. This finalizes the result. □

Lemma B.4 (Maximum Visitation of PseudoSuboptimal Pairs)**.** *With probability at least $1-2\delta$, and for all $\Delta \in [0,1]$ and $T \in \mathbb{N}$ the number of episodes whose sample trajectories contain a state from* $\mathrm{PathPseudoSub}_\Delta$ *is upper bounded by,*

$$\begin{aligned} &\sum_{t=1}^T \mathbf{1}(\tau_t \cap (\mathrm{PathPseudoSub}_\Delta \cup \mathrm{PseudoSub}_\Delta) \neq \emptyset) \\ &\leq |\mathrm{BoundaryPseudoSub}_\Delta| \times \frac{8192\beta^2 \times \left(\widetilde{V}^{\max}\right)^2 \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}{\Delta^2} \end{aligned}$$

and

$$\begin{aligned} &\sum_{(s,a) \in (\{\mathrm{PathPseudoSub}_\Delta \times \mathcal{A}\} \cup \mathrm{PseudoSub}_\Delta)} N_{h(s)}^T(s,a) \\ &\leq H \times |\mathrm{BoundaryPseudoSub}_\Delta| \times \frac{8192\beta^2 \times \left(\widetilde{V}^{\max}\right)^2 \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}{\Delta^2} \end{aligned}$$

Where $h(s)$ corresponds to horizon index of the state partitions that contains state s.

Proof. Let τ_t be the trajectory sampled by our algorithm at time t. Notice that whenever $\tau_t \cap \text{PathPseudoSub}_\Delta \neq \emptyset$, there must be a state action pair in τ_t (occurring previous to the state in τ_t that lies in $\text{PathPseudoSub}_\Delta$) that is in $\text{BoundaryPseudoSub}_\Delta$, namely the first state-action pair belonging to PseudoSub_Δ in τ_t. We obtain,

$$\sum_{t=1}^{T} \mathbf{1}(\tau_t \cap (\text{PathPseudoSub}_\Delta \cup \text{PseudoSub}_\Delta) \neq \emptyset) \leq \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \text{BoundaryPseudoSub}_\Delta \neq \emptyset).$$

Conditioning on the event from Lemma B.3 implies that for all $T \in \mathbb{N}$

$$\begin{aligned}\sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \text{BoundaryPseudoSub}_\Delta \neq \emptyset) &\leq \sum_{(s,a)\in\text{BoundaryPseudoSub}_\Delta} N^T_{h(s)}(s,a) \\ &\leq |\text{BoundaryPseudoSub}_\Delta| \times \frac{8192\beta^2 \times \max_{s',h'} \widetilde{V}^2_{h'}(s') \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{\Delta^2}.\end{aligned}$$

Thus,

$$\begin{aligned}&\sum_{t=1}^{T} \mathbf{1}(\tau_t \cap (\text{PathPseudoSub}_\Delta \cup \text{PseudoSub}_\Delta) \neq \emptyset) \\ &\leq |\text{BoundaryPseudoSub}_\Delta| \times \frac{8192\beta^2 \times \max_{s',h'} \widetilde{V}^2_{h'}(s') \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{\Delta^2}.\end{aligned}$$

Finalizing the proof of the first bullet. To prove the second bullet, for any $t \in \mathbb{N}$ the following inequalities hold

$$\begin{aligned}&\sum_{(s,a)\in(\text{PathPseudoSub}_\Delta\times\mathcal{A}\cup\text{PseudoSub}_\Delta)} N^t_{h(s)}(s,a) \\ &\leq H \sum_{i=1}^{t} \mathbf{1}(\tau_t \cap (\text{PathPseudoSub}_\Delta \times \mathcal{A} \cup \text{PseudoSub}_\Delta) \neq \emptyset) \\ &\leq H \sum_{i=1}^{t} \mathbf{1}(\tau_t \cap \text{BoundaryPseudoSub}_\Delta \neq \emptyset) \\ &\leq H \sum_{(s,a)\in\text{BoundaryPseudoSub}_\Delta} N^t_{h(s)}(s,a) \\ &\leq H \times |\text{BoundaryPseudoSub}_\Delta| \times \frac{8192\beta^2 \times \max_{s',h'} \widetilde{V}^2_{h'}(s') \cdot \ln \frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}}{\Delta^2}.\end{aligned}$$

The result follows. □

Let's now consider a refinement to the upper bound of Equation 3. For any[6] $T \in \mathbb{N}$,

$$\begin{aligned}\sum_{t=1}^{T} V^\star(s_0) - V^{\pi_t}(s_0) &\leq \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^{H} b^t_h(s_h,a_h) + \left(\widehat{\mathbb{P}}_t(\cdot|s_h,a_h) - \mathbb{P}^\star(\cdot|s_h,a_h)\right)\cdot \widehat{V}^{\pi_t}_{h+1}\right] \\ &\leq \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^{H} b^t_h(s_h,a_h) + \left(\widehat{\mathbb{P}}_t(\cdot|s_h,a_h) - \mathbb{P}^\star(\cdot|s_h,a_h)\right)\cdot\left(\widehat{V}^{\pi_t}_{h+1} - V^\star_{h+1}\right)\right]}_{\mathrm{I}} + \\ &\quad \sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^{H}\left(\widehat{\mathbb{P}}_t(\cdot|s_h,a_h) - \mathbb{P}^\star(\cdot|s_h,a_h)\right)\cdot V^\star_{h+1}\right]. \end{aligned} \tag{9}$$

[6] From now on we'll use T to index the final timestep of the regret sequence we are bounding.

By Equation 5 it follows that with probability at least $1-\delta$ for all $s,a,h,T \in \mathcal{S}\times\mathcal{A}\times[H]\times\mathbb{N}$,

$$\sum_{t=1}^{T} \mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^{H}\left(\widehat{\mathbb{P}}_t(\cdot|s_h,a_h) - \mathbb{P}^\star(\cdot|s_h,a_h)\right)\cdot V^\star_{h+1}\right] \le \sum_{t=1}^{T}\mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^{H} b^t_h(s_h,a_h)\right]$$

We will now focus on bounding I. We'll make use of Lemma 7.8 from [2]

Lemma B.5 (Lemma 7.8 from [2])**.** *With probability at least* $1-\delta$ *for all* $t\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}$, *we have,*

$$\left|\widehat{\mathbb{P}}_t(s'|s,a) - \mathbb{P}^\star(s'|s,a)\right| \le \sqrt{\frac{2\mathbb{P}^\star(s'|s,a)\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)}} + \frac{2\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)}$$

Where $h(s)$ *corresponds to horizon index of the state partitions that contains state* s*.*

A slight modification of the proof[7] of Lemma 7.7 in [2] yields,

Lemma B.6 (Support Aware version of Lemma 7.7 in [2])**.** *With probability at least* $1-\delta$ *for all* $t\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}$ *and all* $f:\mathcal{S}\to\mathbb{R}$ *satisfying* $f:\mathcal{S}\to[0,B]$ *we have,*

$$\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right)f\right| \le \frac{\mathbb{E}_{s'\sim\mathbb{P}^\star(\cdot|s,a)}\left[f(s')\right]}{H} + B\min\left(\frac{3\left|\text{support}(\mathbb{P}^\star(\cdot|s,a))\right| H\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)}, 1\right)$$

Where $h(s)$ *corresponds to horizon index of the state partitions that contains state* s*.*

[7] Instead of using $|\mathcal{S}|$ to uniformly bound the support of $\mathbb{P}^\star(\cdot|s,a)$ we explicitly write the bound in terms of its support

Proof. We start by conditioning on the event that Lemma B.5 holds. Take any function $f : \mathcal{S} \to [0,B]$. We have,

$$\begin{aligned}
\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right) f\right| &\le \sum_{s'\in\mathcal{S}} \left|\left(\widehat{\mathbb{P}}_t(s'|s,a) - \mathbb{P}^\star(s'|s,a)\right)\right| f(s') \\
&\le \sum_{s'\in\mathcal{S}} \sqrt{\frac{2\mathbb{P}^\star(s'|s,a)\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right) f^2(s')}{N^t_{h(s)}(s,a)}} + \sum_{s'\in\mathcal{S}} \frac{2\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right) f(s')}{N^t_{h(s)}(s,a)} \\
&\overset{(i)}{\le} \sum_{s'\in\mathcal{S}} \sqrt{\frac{2\mathbb{P}^\star(s'|s,a)\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right) f^2(s')}{N^t_{h(s)}(s,a)}} + \frac{2B\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} \\
&\overset{(ii)}{\le} \sqrt{|\text{support}(\mathbb{P}^\star(\cdot|s,a))|}\sqrt{\frac{\sum_{s'\in\mathcal{S}} 2\mathbb{P}^\star(s'|s,a)\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right) f^2(s')}{N^t_{h(s)}(s,a)}} + \\
&\quad \frac{2B\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} \\
&= \sqrt{\frac{2\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,BH\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} \frac{\sum_{s'\in\mathcal{S}} \mathbb{P}^\star(s'|s,a) f^2(s')}{BH}} + \\
&\quad \frac{2B\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} \\
&\overset{(iii)}{\le} \frac{|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,BH\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} + \frac{\sum_{s'\in\mathcal{S}} \mathbb{P}^\star(s'|s,a) f^2(s')}{BH} + \\
&\quad \frac{2B\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} \\
&\overset{(iv)}{\le} \frac{|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,BH\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} + \frac{\sum_{s'\in\mathcal{S}} \mathbb{P}^\star(s'|s,a) f(s')}{H} + \\
&\quad \frac{2B\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)}
\end{aligned}$$

Inequality (i) follows because $\|f\|_\infty \le B$. Inequality (ii) is a result of Cauchy-Schwarz, (iii) uses the inequality $ab \le \frac{a^2+b^2}{2}$ and (iv) uses the condition $f \in [0,B]$ again. The above display implies,

$$\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right) f\right| \le \frac{3\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,BH\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)} + \frac{\sum_{s'\in\mathcal{S}} \mathbb{P}^\star(s'|s,a) f(s')}{H}.$$

Finally, since $f(s) \in [0,B]$,

$$\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right) f\right| \le B.$$

Combining these last two equations we conclude,

$$\left|\left(\widehat{\mathbb{P}}_t(\cdot|s,a) - \mathbb{P}^\star(\cdot|s,a)\right) f\right| \le \min\left(B, \frac{3\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,BH\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N^t_{h(s)}(s,a)}\right) + \frac{\sum_{s'\in\mathcal{S}} \mathbb{P}^\star(s'|s,a) f(s')}{H}.$$

The result follows. □

From now on we'll use the notation

$$\xi_h^t(s,a) := (\beta-1)\widetilde{V}^{\max}\min\left(\frac{3\,|\text{support}(\mathbb{P}^\star(\cdot|s,a))|\,H\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N_h^t(s,a)},1\right).$$

Recall that by Assumption 3, if we define $h(s)$ as to horizon index of the state partitions that contains state s, we have $N_h(s,a)=0$ for all $h\neq h(s)$.

The following corollary holds,

Corollary B.7. *With probability at least $1-2\delta$ for all $t\in\mathbb{N}, s\in\mathcal{S}, a\in\mathcal{A}, h\in[H]$ we have,*

$$\left(\widehat{\mathbb{P}}_t(\cdot|s,a)-\mathbb{P}^\star(\cdot|s,a)\right)\left(\widehat{V}_{h+1}^{\pi_t}-V_{h+1}^\star\right)\leq\frac{\mathbb{E}_{s'\sim\mathbb{P}^\star(\cdot|s,a)}\left[\widehat{V}_{h+1}^{\pi_t}(s')-V_{h+1}^\star(s')\right]}{H}+\xi_h^t(s,a)$$

Proof. Under the event of Lemma 5.1 (optimism), it follows that $\widehat{V}_{h+1}^{\pi_t}(s')\geq V_{h+1}^\star(s')$ for all $t\in\mathbb{N}, s\in\mathcal{S}$. Moreover, $\widehat{V}_{h+1}^{\pi_t}(s')-V_{h+1}^\star(s')\leq\beta\widetilde{V}_{h+1}(s')-V_{h+1}^\star(s')\leq(\beta-1)\widetilde{V}_{h+1}(s')$.

Thus we can set $B=(\beta-1)\widetilde{V}^{\max}$ in Lemma B.6. □

We will now restate a modified version of Lemma 7.10 from [2] that provides us a bound for term I in terms of expectations over sums of b_h^t and ξ_h^t terms.

Lemma B.8 (Notation Adapted version of Lemma 7.10 from [2])**.** *For all $T\in\mathbb{N}$ with probability at least $1-3\delta$,*

$$\mathrm{I}\leq e\sum_{t=1}^T\mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^H b_h^t(s_h,a_h)+\xi_h^t(s_h,a_h)\right].$$

Thus, with probability at least $1-3\delta$,

$$\sum_{t=1}^T V^\star(s_0)-V^{\pi_t}(s_0)\leq e\sum_{t=1}^T\mathbb{E}_{\tau\sim\pi_t}\left[\sum_{h=1}^H 2b_h^t(s_h,a_h)+\xi_h^t(s_h,a_h)\right].$$

The following Lemma will allow us to change from a sum of expectations over the played policies to a sum over the sample trajectories.

Lemma B.9. *Simultaneously for all $\mathcal{U}\subset\mathcal{S}\times\mathcal{A}$ and all $T\in\mathbb{N}$.*

$$\begin{aligned}\sum_{t=1}^T V^\star(s_0)-V^{\pi_t}(s_0)\leq{}& e\sum_{t=1}^T\sum_{h=1}^H\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U})\left(2b_h^t(s_h^t,a_h^t)+\xi_h^t(s_h^t,a_h^t)\right)+\\&e\sum_{t=1}^T\sum_{h=1}^H\mathbf{1}((s_h^t,a_h^t)\notin\mathcal{U})\left(2b_h^t(s_h^t,a_h^t)+\xi_h^t(s_h^t,a_h^t)\right)+\\&\mathcal{O}\left(\beta\widetilde{V}^{\max}\sqrt{HT\ln\left(\frac{T}{\delta}\right)}\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{t=1}^T V^\star(s_0)-V^{\pi_t}(s_0)\leq{}& e\sum_{t=1}^T\mathbf{1}(\tau_t\cap\mathcal{U}=\emptyset)\sum_{h=1}^H\left(2b_h^t(s_h^t,a_h^t)+\xi_h^t(s_h^t,a_h^t)\right)+\\&\mathcal{O}\left(\beta\widetilde{V}^{\max}\sqrt{HT\ln\left(\frac{T}{\delta}\right)}+\beta H\widetilde{V}^{\max}\sum_{t=1}^T\mathbf{1}(\tau_t\cap\mathcal{U}\neq\emptyset)\right)\end{aligned}$$

With probability at least $1-4\delta$.

Proof. By Lemma B.8, for all $T \in \mathbb{N}$ with probability at least $1-3\delta$,

$$\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) \le e \sum_{t=1}^{T} \mathbb{E}_{\tau \sim \pi_t}\left[\sum_{h=1}^{H} 2b_h^t(s_h, a_h) + \xi_h^t(s_h, a_h)\right]. \tag{10}$$

Since $|2b_h^t(s_h,a_h) + \xi_h^t(s_h,a_h)| \le \mathcal{O}\left(\beta \widetilde{V}^{\max}\right)$, anytime Hoeffding lemma C.1 and the union bound implies that for all $T \in \mathbb{N}$ with probability at least $1-4\delta$,

$$\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) \le e \sum_{t=1}^{T}\sum_{h=1}^{H} 2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t) + \mathcal{O}\left(\beta \widetilde{V}^{\max}\sqrt{HT\ln\left(\frac{T}{\delta}\right)}\right) \tag{11}$$

Observe that for any $\mathcal{U} \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H} 2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t) &= \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} = \emptyset) \sum_{h=1}^{H} \left(2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t)\right) + \\ &\quad \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} \neq \emptyset) \sum_{h=1}^{H} \left(2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t)\right) \\ &\le \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} = \emptyset) \sum_{h=1}^{H} \left(2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t)\right) + \\ &\quad \mathcal{O}\left(\beta H \widetilde{V}^{\max} \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} \neq \emptyset)\right). \end{aligned} \tag{12}$$

The result follows by combining inequalities 10, 11 and 12.

□

B.1 Full Proof of Theorem 5.2

We are ready to prove our main supporting Lemma,

Lemma B.10. *With probability at east $1-6\delta$ for all $\Delta > 0$ and $T \in \mathbb{N}$ simultaneously,*

$$\begin{aligned}\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) \leq \beta H \widetilde{V}^{\max} \sqrt{|\mathcal{U}^{\text{good}}| T \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \\ \mathcal{O}\Bigg(\min\Bigg((\beta-1)\widetilde{V}^{\max} |\mathcal{S}|^2 |\mathcal{A}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1) + \\ \beta H \widetilde{V}^{\max} \sqrt{|\mathcal{U}^{\text{good}}| T \ln \frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}} + \\ \beta \left(\widetilde{V}^{\max}\right)^2 \left(\ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right) \frac{\sqrt{H|\mathcal{S}||\mathcal{A}||\text{BoundaryPseudoSub}_{\Delta}|}}{\Delta} + \\ \beta \widetilde{V}^{\max} |\mathcal{S}||\mathcal{A}| \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta} \ln(T+1), \\ (\beta-1)\widetilde{V}^{\max} |\mathcal{S}\backslash \text{PathPseudoSub}_{\Delta}| \, |\mathcal{U}^{\text{good}}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1) + \\ \beta \widetilde{V}^{\max} |\mathcal{U}^{\text{good}}| \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta} \ln(T+1) + \\ \beta H \widetilde{V}^{\max} \sqrt{|\mathcal{U}^{\text{good}}| T \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}} + \\ \beta^3 H \left(\widetilde{V}^{\max}\right)^3 |\text{BoundaryPseudoSub}_{\Delta}| \times \frac{\ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\Delta^2} \Bigg)\Bigg)\end{aligned}$$

Where $\mathcal{U}^{\text{good}} = (\mathcal{S} \times \mathcal{A}) \backslash (\text{PathPseudoSub}_{\Delta} \times \mathcal{A} \cup \text{PseudoSub}_{\Delta})$.

Proof. We will condition on the events from Lemmas B.4 and B.9, something that happens with probability at least $1 - 6\delta$.

We'll start with the decomposition from Lemma B.9. Set $\mathcal{U}^{\text{good}} = (\mathcal{S} \times \mathcal{A}) \backslash [(\text{PathPseudoSub}_{\Delta} \times \mathcal{A}) \cup \text{PseudoSub}_{\Delta}]$. In what follows we'll use $\mathcal{U}^{\text{good}}$ to denote this set when convenient. We will also use the notation $\mathcal{U}^{\text{bad}} = (\mathcal{S} \times \mathcal{A}) \backslash \mathcal{U}^{\text{good}}$ to denote the complement of $\mathcal{U}_{\text{good}}$.

Recall that as a consequence of Equation 4 for all state action pairs $(s,a) \in \mathcal{U}^{\text{good}}$, we have that $\text{Neighbor}(s,a) \subseteq \mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}$.

Therefore, the support of $\mathbb{P}^{\star}(\cdot|s,a)$ and $\widehat{\mathbb{P}}_t(\cdot|s,a)$ is contained in $\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}$ for $(s,a) \in \mathcal{S} \backslash (\text{PathPseudoSub}_{\Delta} \times \mathcal{A} \cup \text{PseudoSub}_{\Delta})$. This implies that for all $(s,a) \in (\mathcal{S} \times \mathcal{A}) \backslash (\text{PathPseudoSub}_{\Delta} \times \mathcal{A} \cup \text{PseudoSub}_{\Delta})$, we have

$$\xi_h^t(s,a) \leq (\beta-1)\widetilde{V}^{\max} \min\left(\frac{3\,|\mathcal{S}\backslash \text{PathPseudoSub}_{\Delta}|\, H \ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N_h^t(s,a)}, 1 \right) \;\forall\, (s,a) \in \mathcal{U}^{\text{good}}. \tag{13}$$

Similarly,

$$\xi_h^t(s,a) \leq (\beta-1)\widetilde{V}^{\max} \min\left(\frac{3\,|\mathcal{S}|\, H \ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N_h^t(s,a)}, 1 \right) \;\forall\, (s,a) \in \mathcal{U}^{\text{bad}}. \tag{14}$$

Let $\mathcal{U}^{\text{good}} = (\mathcal{S} \times \mathcal{A}) \backslash (\text{PathPseudoSub}_\Delta \times \mathcal{A} \cup \text{PseudoSub}_\Delta)$. Inequality 13 implies,

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) \xi_h^t(s_h^t, a_h^t) && (15)\\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) (\beta - 1) \widetilde{V}^{\max} \min\left(\frac{3\,|\mathcal{S} \backslash \text{PathPseudoSub}_\Delta|\, H \ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N_h^t(s_h^t, a_h^t)}, 1 \right)\\
&\leq 3(\beta - 1)\widetilde{V}^{\max}\,|\mathcal{S} \backslash \text{PathPseudoSub}_\Delta|\, H \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}})}{N_h^t(s_h^t, a_h^t)}\\
&\overset{(i)}{\leq} 6(\beta - 1)\widetilde{V}^{\max}\,|\mathcal{S} \backslash \text{PathPseudoSub}_\Delta|\, |\mathcal{U}^{\text{good}}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1).
\end{aligned}$$

Where inequality (i) holds because of Lemma C.5. The exact same argument applied to the upper bound of Equation 15 implies,

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{bad}}) \xi_h^t(s_h^t, a_h^t) &\leq 6(\beta - 1)\widetilde{V}^{\max}\,|\mathcal{S}|\, |\mathcal{U}^{\text{bad}}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1)\\
&\leq 6(\beta - 1)\widetilde{V}^{\max}\,|\mathcal{S}|^2\, |\mathcal{A}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1). \quad (16)
\end{aligned}$$

Let's now bound the sum $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) b_h^t(s_h^t, a_h^t)$.

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) b_h^t(s_h^t, a_h^t) && (17)\\
&= \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) \min\left(16\beta \sqrt{\frac{\widehat{\mathbb{E}}_{s' \sim \widehat{\mathbb{P}}_t(\cdot|s_h^t, a_h^t)}[\widetilde{V}_{h+1}^2(s') | s_h^t, a_h^t] \ln \frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}{N_h^t(s_h^t, a_h^t)}} + \frac{12\beta\widetilde{V}^{\max}}{N_h^t(s_h^t, a_h^t)} \ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta}, 2\beta\widetilde{V}^{\max} \right)\\
&\leq \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) \left(16\beta \sqrt{\frac{\widehat{\mathbb{E}}_{s' \sim \widehat{\mathbb{P}}_t(\cdot|s_h^t, a_h^t)}[\widetilde{V}_{h+1}^2(s') | s_h^t, a_h^t] \ln \frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}{N_h^t(s_h^t, a_h^t)}} + \frac{12\beta\widetilde{V}^{\max}}{N_h^t(s_h^t, a_h^t)} \ln\frac{2|\mathcal{S}||\mathcal{A}|t}{\delta} \right)\\
&\overset{(a)}{\leq} 16\beta\widetilde{V}^{\max}\sqrt{\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}} \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}})}{\sqrt{N_h^t(s_h^t, a_h^t)}} + 12\beta\widetilde{V}^{\max} \ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta} \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}})}{N_h^t(s_h^t, a_h^t)}\\
&\overset{(i)}{\leq} 32\beta H\widetilde{V}^{\max}\sqrt{|\mathcal{U}^{\text{good}}| T \ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}} + 24\beta\widetilde{V}^{\max}|\mathcal{U}^{\text{good}}| \ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta} \ln(T+1).
\end{aligned}$$

Where inequality (i) holds because of Lemmas C.4 and C.5. The exact same proof argument as in the proof of inequality 17 that leads to inequality (a) yields,

$$
\begin{aligned}
&\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U}^{\text{bad}})b_h^t(s_h^t,a_h^t) \qquad (18)\\
&\le 16\beta\widetilde{V}^{\max}\sqrt{\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U}^{\text{bad}})}{\sqrt{N_h^t(s_h^t,a_h^t)}}+12\beta\widetilde{V}^{\max}\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U}^{\text{bad}})}{N_h^t(s_h^t,a_h^t)}\\
&\overset{(i)}{\le} 32\beta\widetilde{V}^{\max}\sqrt{H|\mathcal{U}^{\text{bad}}|\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\sum_{(s,a)\in(\mathcal{S}\times\mathcal{A})\setminus\mathcal{U}^{\text{good}}}\sum_{h=1}^{H}N_h^T(s,a)}+24\beta\widetilde{V}^{\max}|\mathcal{U}^{\text{bad}}|\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\\
&\overset{(ii)}{\le} 32\times 8192\beta\left(\widetilde{V}^{\max}\right)^2\left(\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\right)\frac{\sqrt{H|\mathcal{U}^{\text{bad}}||\text{BoundaryPseudoSub}_{\Delta}|}}{\Delta}+\\
&\quad 24\beta\widetilde{V}^{\max}|\mathcal{U}^{\text{bad}}|\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\\
&=\mathcal{O}\left(\beta\left(\widetilde{V}^{\max}\right)^2\left(\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\right)\frac{\sqrt{H|\mathcal{S}||\mathcal{A}||\text{BoundaryPseudoSub}_{\Delta}|}}{\Delta}\right)+\\
&\quad\mathcal{O}\left(\beta\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\right),
\end{aligned}
$$

where inequality (i) follows because of Lemmas C.4 and C.5. And inequality (ii) follows from Lemma B.4.

Combining inequalities 15, 16, 17 and 18 with Lemma B.9 we get,

$$
\begin{aligned}
\sum_{t=1}^{T}V^{\star}(s_0)-V^{\pi_t}(s_0) &\le e\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U}^{\text{good}})\left(2b_h^t(s_h^t,a_h^t)+\xi_h^t(s_h^t,a_h^t)\right)+\\
&\quad e\sum_{t=1}^{T}\sum_{h=1}^{H}\mathbf{1}((s_h^t,a_h^t)\in\mathcal{U}^{\text{bad}})\left(2b_h^t(s_h^t,a_h^t)+\xi_h^t(s_h^t,a_h^t)\right)+\mathcal{O}\left(\beta\widetilde{V}^{\max}\sqrt{HT\ln\left(\frac{T}{\delta}\right)}\right)\\
&\le\mathcal{O}\left((\beta-1)\widetilde{V}^{\max}|\mathcal{S}|^2|\mathcal{A}|H^2\ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right)\log(T+1)\right)+\\
&\quad\mathcal{O}\left(\beta H\widetilde{V}^{\max}\sqrt{|\mathcal{U}^{\text{good}}|T\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}\right)+\\
&\quad\mathcal{O}\left(\beta\left(\widetilde{V}^{\max}\right)^2\left(\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\right)\frac{\sqrt{H|\mathcal{S}||\mathcal{A}||\text{BoundaryPseudoSub}_{\Delta}|}}{\Delta}\right)+\\
&\quad\mathcal{O}\left(\beta\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\right)
\end{aligned}
$$

We will now derive a bound that has only logarithmic dependence on $\mathcal{S}$ at the cost of a quadratic dependence on Δ^2.

By Lemma B.9, inequalities 15 and 17 and Lemma B.4.

$$
\begin{aligned}
\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) &\le e\sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U}^{\text{bad}} = \emptyset) \sum_{h=1}^{H} \left(2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t)\right) + \\
&\quad \mathcal{O}\left(\beta \widetilde{V}^{\max} \sqrt{HT \ln\left(\frac{T}{\delta}\right)} + \beta H \widetilde{V}^{\max} \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} \neq \emptyset)\right) \\
&\le e \sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}((s_h^t, a_h^t) \in \mathcal{U}^{\text{good}}) \left(2b_h^t(s_h^t, a_h^t) + \xi_h^t(s_h^t, a_h^t)\right) + \\
&\quad \mathcal{O}\left(\beta \widetilde{V}^{\max} \sqrt{HT \ln\left(\frac{T}{\delta}\right)} + \beta H \widetilde{V}^{\max} \sum_{t=1}^{T} \mathbf{1}(\tau_t \cap \mathcal{U} \neq \emptyset)\right) \\
&\overset{(i)}{=} \mathcal{O}\left((\beta - 1)\widetilde{V}^{\max} \left|\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}\right| |\mathcal{U}^{\text{good}}| H^2 \ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right) \log(T+1)\right) + \\
&\quad \mathcal{O}\left(\beta \widetilde{V}^{\max} |\mathcal{U}^{\text{good}}| \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta} \ln(T+1)\right) + \\
&\quad \mathcal{O}\left(\beta H \widetilde{V}^{\max} \sqrt{|\mathcal{U}^{\text{good}}| T \ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}\right) + \\
&\quad \mathcal{O}\left(\beta^3 H \left(\widetilde{V}^{\max}\right)^3 |\text{BoundaryPseudoSub}_{\Delta}| \times \frac{\ln \frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}{\Delta^2}\right)
\end{aligned}
$$

This holds for all $\Delta \in [0,1]$ simultaneously because Lemma B.4 holds simultaneously for all $\Delta \in [0,1]$ and for all $T \in \mathbb{N}$ because all the events we have conditioned refer to properties that hold for all $T \in \mathbb{N}$. This finalizes the proof of the desired result. □

Using the bound $|\mathcal{U}^{\text{good}}| \le \left|\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}\right| |\mathcal{A}|$, we have the following version of our results

Theorem B.11 (Full version of Theorem 5.2)**.** *The regret of UCBVI - Shaped satisfies,*

$$
\begin{aligned}
\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) = \mathcal{O}\Bigg(\min_{\Delta} \Bigg(& H\beta\widetilde{V}^{\max} \sqrt{|\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}||\mathcal{A}|T \ln \frac{\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|T}{\delta}} + \\
& \beta\widetilde{V}^{\max} \ln \frac{\widetilde{V}^{\max}|\mathcal{S}||\mathcal{A}|T}{\delta} \times \min\left(\bar{A}(\Delta), \bar{B}(\Delta)\right) \Bigg)\Bigg).
\end{aligned}
$$

For all $T \in \mathbb{N}$ with probability at least $1 - 6\delta$. Where

$$
\begin{aligned}
\bar{A}(\Delta) = & \frac{\beta\widetilde{V}^{\max} H^{1/2} |\mathcal{S}|^{1/2} |\mathcal{A}|^{1/2} |\text{BoundaryPseudoSub}_{\Delta}|^{1/2}}{\Delta} + \frac{(\beta - 1)}{\beta} |\mathcal{S}|^2 |\mathcal{A}| H^2 \log(T+1) \\
& + |\mathcal{S}||\mathcal{A}| \ln(T+1),
\end{aligned}
$$

and

$$
\begin{aligned}
\bar{B}(\Delta) = & \frac{\beta^2 \left(\widetilde{V}^{\max}\right)^2 H |\text{BoundaryPseudoSub}_{\Delta}|}{\Delta^2} + \frac{(\beta - 1)}{\beta} \left|\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}\right|^2 |\mathcal{A}| H^2 \log(T+1) \\
& + \left|\mathcal{S} \backslash \text{PathPseudoSub}_{\Delta}\right| |\mathcal{A}| \ln(T+1).
\end{aligned}
$$

B.2 Generalizations

In this section we briefly discuss what can we say in the case the shaping functions $\{\widetilde{Q}_h\}_{h\in[H]}$ are of the form $\widetilde{Q}_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that $Q_h^{\star}(s,a) \le \beta \widetilde{Q}_h(s,a)$ instead of our state-only assumption. By doing so we recover some of the results of [21] for which we provide simple proofs. We will use the notation $\widetilde{Q}^{\max} = \max_{s,a,h} \widetilde{Q}_h(s,a)$.

Let's consider a generic optimistic RL algorithm $\mathbb{A}$ that computes an *optimistic Q* function estimator $\{\widehat{Q}_h^t\}_{h\in[H]}$ at the beginning of time-step t satisfying $\widehat{Q}_h^t(s,\pi_\star(s)) \geq Q_\star(s,\pi_\star(s))$. Notice that we only consider optimism at the optimal action $\pi_\star(s)$. Let's also assume the policy played by $\mathbb{A}$ at time t equals $\pi_t(|s,a,h) = \arg\max_{a\in\mathcal{A}} \widehat{Q}_h^t(s,a)$. We will introduce the notion of a least upper bound stability for optimistic algorithms.

Definition 1. *We say that an optimistic algorithm $\mathbb{A}$ is stable w.r.t. to clipping if substituting the $\{\widetilde{Q}_h^t\}_{h\in[H]}$ values by their clipped versions*

$$\widehat{Q}_h^{t,\text{clipped}}(s,a) \leftarrow \min\left(\widehat{Q}_h^t(s,a), \beta\widetilde{Q}(s,a)\right)$$

does not affect its regret guarantees.

Clipping is done *after* the $\widehat{Q}_h^t(s,a)$ values have been computed.

Optimism is preserved. When the un-clipped $\{\widehat{Q}_h^t\}_{h\in[H]}$ are optimistic, then as long as $\beta\widetilde{Q}_h(s,\pi_\star(s)) \geq Q_\star(s,\pi_\star(s))$, the maximum among all the clipped values is also optimistic.

$$\max_{a\in\mathcal{A}} \widehat{Q}_h^{t,\text{clipped}}(s,a) = \max_{a\in\mathcal{A}}\left[\min\left(\widehat{Q}_h^t(s,a), \beta\widetilde{Q}(s,a)\right)\right] \geq \min\left(\widehat{Q}_h^t(s,\pi_\star(s)), \beta\widetilde{Q}(s,\pi_\star(s))\right) \geq Q^\star(s,\pi_\star(s))$$

If $\mathbb{A}$'s policy at time t equals $\pi_t(|s,h) = \arg\max_{a\in\mathcal{A}} \widehat{Q}_h^{t,\text{clipped}}(s,a)$, as long as the un-clipped $\{\widehat{Q}_h^t\}_{h\in[H]}$ are optimistic the support of policy π_t satisfies

$$\text{Support}(\pi_t(\cdot|s,a,h)) = \{a' \text{ s.t. } \beta\widetilde{Q}_h(s,a') \geq Q_h^\star(s,\pi_\star(s,h))\}$$

This is simply because for all $a \in \mathcal{A}\backslash\{a' \text{ s.t. } \beta\widetilde{Q}_h(s,a') \geq Q_h(s,\pi_\star(s,h)\}$, the clipped $\{\widehat{Q}_h^{t,\text{clipped}}(s,a)\}_{h\in[H]}$ satisfy

$$\widehat{Q}_h^{t,\text{clipped}}(s,a) < Q_h^\star(s,\pi_\star(s,h))$$

In other words, for all $t \in \mathbb{N}$ the clipped algorithm will only visit state-action pairs in $\cup_{s\in\mathcal{S},h\in[H]}\{a \text{ s.t. } \beta\widetilde{Q}_h(s,a) \geq Q_h^\star(s,\pi_\star(s,h))\}$. Alternatively this can be thought of as if the learner were to interact only with a prunned MDP with state space $\mathcal{S}$ and state (and horizon) dependent action sets $\mathcal{A}(s)$ of the form $\mathcal{A}(s) = \{a \text{ s.t. } \beta\widetilde{Q}_{h(s)}(s,a) \geq Q_{h(s)}^\star(s,\pi_\star(s,h(s)))\}$.

It is easy to see the corresponding version of UCBVI-Shaped satisfies the following regret guarantee that recovers the min-max rates of Theorem 1.1 in [21],

Lemma B.12. *The regret of UCBVI-Shaped with $\widetilde{Q}$-shaping satisfies,*

$$\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0) \leq \mathcal{O}\left(\beta\widetilde{Q}^{\max}\sqrt{|\text{Pair}_{\text{eff}}|T\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}\right) + \mathcal{O}\left(\beta\widetilde{Q}^{\max}H|\mathcal{S}|\,|\text{Pair}_{\text{eff}}|\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\right)$$

For all $T \in \mathbb{N}$ with probability at least $1-\delta$, where[8] $\text{Pair}_{\text{eff}} = \{(s,a) \in \mathcal{S}\times\mathcal{A}$ *s.t.* $a \in \mathcal{A}(s)\}$.

Proof. We borrow the bound from Theorem 7.6 in [2]. Observe that as long as optimism holds (Lemma 5.1), the regret decomposition of Lemma B.9 is satisfied.

$$\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0) \leq e\sum_{t=1}^T\sum_{h=1}^H \left(2b_h^t(s_h^t,a_h^t) + \xi_h^t(s_h^t,a_h^t)\right) + \mathcal{O}\left(\beta\widetilde{Q}^{\max}\sqrt{HT\ln\left(\frac{T}{\delta}\right)}\right)$$

Notice that as long as optimism holds, for all s_h^t the action a_h^t satisfies $a_h^t \in \mathcal{A}(s_h^t)$. Thus,

[8] Recall we have assumed that $\mathcal{S}$ is h indexed so that any state $s \in \mathcal{S}$ can be accessed only during in-episode time-step $h(s)$.

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\xi_h^t(s_h^t,a_h^t) \le \sum_{t=1}^{T}\sum_{h=1}^{H}\beta\widetilde{Q}^{\max}\min\left(\frac{3\,|\mathcal{S}|\,H\ln\left(\frac{|\mathcal{S}||\mathcal{A}|tH}{\delta}\right)}{N_h^t(s_h^t,a_h^t)},1\right)$$
$$\le 3\beta\widetilde{Q}^{\max}\,|\mathcal{S}|\,H\ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right)\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{N_h^t(s_h^t,a_h^t)}$$
$$\overset{(i)}{\le} 6\beta\widetilde{Q}^{\max}\,|\mathcal{S}|\,\mathrm{Pair}_{\mathrm{eff}}\cdot H\ln\left(\frac{|\mathcal{S}||\mathcal{A}|TH}{\delta}\right)\log(T+1).$$

Where inequality (i) follows from Lemma C.5. Similarly,

$$\sum_{t=1}^{T}\sum_{h=1}^{H}b_h^t(s_h^t,a_h^t) \le 16\beta\widetilde{Q}^{\max}\sqrt{\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{\sqrt{N_h^t(s_h^t,a_h^t)}}$$
$$+12\beta\widetilde{Q}^{\max}\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{1}{N_h^t(s_h^t,a_h^t)}$$
$$\overset{(i)}{\le} 32\beta\widetilde{Q}^{\max}\sqrt{|\mathrm{Pair}_{\mathrm{eff}}|T\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}}$$
$$+24\beta\widetilde{Q}^{\max}|\mathrm{Pair}_{\mathrm{eff}}|\ln\frac{2|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)$$

Where (i) holds by Lemmas C.4 and C.5. Combining these last two inequalities yields

$$\sum_{t=1}^{T}V^{\star}(s_0)-V^{\pi_t}(s_0) \le \mathcal{O}\left(\beta\widetilde{Q}^{\max}\sqrt{|\mathrm{Pair}_{\mathrm{eff}}|T\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}}\right)+$$
$$\mathcal{O}\left(\beta\widetilde{Q}^{\max}H\,|\mathcal{S}|\,|\mathrm{Pair}_{\mathrm{eff}}|\ln\frac{|\mathcal{S}||\mathcal{A}|T}{\delta}\ln(T+1)\right)$$

The result follows.

□

Similarly, we can obtain a result for instance dependent bounds for a clipped version of the StrongEuler algorithm. We consider the exact same clipping mechanism as in the previous discussion. Optimism guarantees (just as we have described above) that $\mathrm{Support}(\pi_t(\cdot|s,a,h)) = \{a' \text{ s.t. } \beta\widetilde{Q}_h(s,a') \ge Q_h^{\star}(s,\pi_{\star}(s,h))\}$. Therefore the data generated by the clipped StrongEuler algorithm will be produced from the 'prunned' MDP with state space $\mathcal{S}$ and state-horizon dependent action sets $\mathcal{A}(s)$ (i.e. with state-action space $\mathrm{Pair}_{\mathrm{eff}}$). Strong optimism holds for the estimates $\widehat{Q}_h^t(s,a)$ for all $(s,a)\in\mathrm{Pair}_{\mathrm{eff}}$ and all time steps t (notice that π_* is an optimal policy both in the original as well as in the pruned MDP). The exact same proofs as in [16] (see proof of Theorem 3.2) hold in this case with the only modification being to adapt the argument for MDPs where the number of actions is state-dependent thus yielding the following result,

Lemma B.13. *The regret of clipped-StrongEuler satisfies with high probability,*

$$\sum_{t=1}^{T}V^{\star}(s_0)-V^{\pi_t}(s_0) = \mathcal{O}\left(\sum_{(s,a)\in\mathrm{Pair}_{\mathrm{eff}}}\frac{Q^{\star}_{h(s)}(s,a)}{\mathrm{gap}_{h(s)}(s,a)}\log(T)\right)$$

where $\mathrm{gap}_h(s,a) = V_h^{\star}(s)-Q_h^{\star}(s,a)$

Stating this result in terms of the return gaps (see Definition 3.1 in [16]) over $\mathrm{Pair}_{\mathrm{eff}}$ is possible. Lemma B.13 recovers the instance dependent results of [21].

B.3 Connections with instance dependent rates for reinforcement learning

Although our main results are not instance dependent, our rates can be much better if the gaps over the effective state space are very small and the effective state space is also very small. Moreover, in contrast with instance dependent rates our results depend on an effective state space size that may be much smaller than the original state space. Our main innovation lies in these state pruning results.

B.4 Model Selection

Although UCBVI-Shaped requires knowledge of β, Algorithm 1 can be used to design a parameter free version. As explained in Section 6 we simply require to initialize algorithm 1 with a set of N different β parameter guesses $[\beta_1, \cdots, \beta_N]$. A good choice for these is an exponential parameter grid of the form $\beta_1 = 1, \cdots, \beta_N = 2^N$.

In the case we are using the Corral algorithm of [34] (and [1]), it is enough to set the learning rates based on a putative regret bound value of $\sqrt{T}$ (see the proof of Theorem 5.3 in [34] where the learning rate can be set to $\eta = \sqrt{\frac{N}{T}}$). The main term in the regret guarantee of Theorem 5.2 scales with $\sqrt{T}$. Thus, for any choice of β_i if this value of β_i was valid, the regret rate would scale as $C_i\sqrt{T\log(1/\delta)}$ for some parameter $C_i \in [1, H\beta_i\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|}]$. If using regret Balancing as an online model selection mechanism (see [33]) we can use an exponential parameter grid of the form $\left\{\beta_i \times [1, 2, \cdots, H\beta_i\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|}\right\}_{i=1}^N$. The number of base algorithms necessary in this case is at most $N\log(H\beta_N\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|})$.

As a consequence of Theorem 5.3 and as a simple corollary of Theorem 5.3 in [34] or Theorem 5.1 in [33] we conclude that,

Lemma B.14. *Provided that $\beta_N\widetilde{V}_h(s) \geq V_h^\star(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ (although it could be that $\beta_i\widetilde{V}_h(s) \geq V_h^\star(s)$ for all $s \in \mathcal{S}$ and $h \in [H]$ for $i \ll N$), the expected regret of Algorithm 1 satisfies,*

$$\mathbb{E}\left[\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0)\right] = \widetilde{\mathcal{O}}\left(C^2\sqrt{NT}\right)$$

when using Stochastic CORRAL as a model selection strategy. Similarly the expected regret of Algorithm 1 satisfies,

$$\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0) = \widetilde{\mathcal{O}}\left(C^2\sqrt{N\log(H\beta_N\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|})T\log(1/\delta)}\right)$$

with probability at least $1 - \mathcal{O}(\delta)$ when using Regret Balancing as a model selection strategy.

Proof. This result is (almost) an immediate corollary of Theorems 5.3 and 5.1 in [34] and [34] respectively. In order to apply these results to our setting we only need to note that for the optimal value of β (among the choices in $[\beta_1, \cdots, \beta_N]$) there exists a constant $C \in [1, 2, \cdots, H\beta_N\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|}]$ such that regret rate of UCBVI-Shaped (Theorem 5.2) can be upper bounded as,

$$\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0) = \widetilde{\mathcal{O}}\left(C\sqrt{T\log(1/\delta)}\right).$$

with probability at least $1 - \mathcal{O}(\delta)$ and where $\widetilde{\mathcal{O}}$ may hide polynomial factors in $|\mathcal{S}|$ and $|\mathcal{A}|$.

Applying the in-expectation results of Theorem 5.3 in [34] for the Stochastic CORRAL model selection algorithm yields (setting $\delta \ll \frac{1}{T}$),

$$\mathbb{E}\left[\sum_{t=1}^T V^\star(s_0) - V^{\pi_t}(s_0)\right] = \widetilde{\mathcal{O}}\left(C^2\sqrt{NT\log(1/\delta)}\right)$$

Where $\widetilde{\mathcal{O}}$ may hide polynomial factors in $|\mathcal{S}|$ and $|\mathcal{A}|$ and T. Similarly, Regret Balancing yields a bound,

$$\sum_{t=1}^{T} V^{\star}(s_0) - V^{\pi_t}(s_0) = \widetilde{\mathcal{O}}\left(C^2\sqrt{N\log(H\beta_N\widetilde{V}^{\max}\sqrt{|\mathcal{S}||\mathcal{A}|})T\log(1/\delta)}\right)$$

with probability at least $1-\mathcal{O}(\delta)$.

□

The pseudocode for this algoritihm is provided below:

Algorithm 3 Online UCBVI - Shaped

```
Input reward function $r$ (assumed to be known), set of $\beta$ values - $[\beta_1, \beta_2, \ldots, \beta_N]$
Initialize model selection probability $p(\beta)$ as a uniform distribution over $[\beta_1, \beta_2, \ldots, \beta_N]$
for $t = 1, ..., T$
    Sample a value of $\beta \sim p(\beta)$
    Run UCBVI-Shaped ($\beta$) with this sampled beta
    Update $p(\beta)$ using samples from UCBVI-Shaped using an online model selection algorithm.
End for
```

C Supporting Technical Lemmas

Lemma C.1 (Anytime Hoeffding Inequality [33]). *Let $\{Y_\ell\}_{\ell=1}^{\infty}$ be a martingale difference sequence such that Y_ℓ is $Y_\ell \in [a_\ell, b_\ell]$ almost surely for some constants a_ℓ, b_ℓ almost surely for all $\ell = 1, \cdots, t$. then*

$$\sum_{\ell=1}^{t} Y_\ell \leq 2\sqrt{\sum_{\ell=1}^{t}(b_\ell - a_\ell)^2 \ln\left(\frac{12t^2}{\widetilde{\delta}}\right)}$$

with probability at least $1-\widetilde{\delta}$ for all $t \in \mathbb{N}$ simultaneously.

Lemma C.2 (Freedman Bounded RVs - unsimplified [19, 8]). *Suppose $\{X_t\}_{t=1}^{\infty}$ is a martingale difference sequence with $|X_t| \leq b$. Let*

$$\mathrm{Var}_\ell(X_\ell) = \mathrm{Var}(X_\ell | X_1, \cdots, X_{\ell-1})$$

Let $V_t = \sum_{\ell=1}^{t} \mathrm{Var}_\ell(X_\ell)$ be the sum of conditional variances of X_t. Then we have that for any $\widetilde{\delta} \in (0,1)$ and $t \in \mathbb{N}$

$$\mathbb{P}\left(\sum_{\ell=1}^{t} X_\ell > 2\sqrt{V_t}A_t + 3bA_t^2\right) \leq \widetilde{\delta}$$

Where $A_t = \sqrt{2\ln\ln\left(2\left(\max\left(\frac{V_t}{b^2}, 1\right)\right)\right) + \ln\frac{6}{\delta}}$ and $h(s)$ corresponds to horizon index of the state partitions that contains state s.

Lemma C.3 (Empirical Bernstein Anytime, Theorem 4 of [29]). *Let $\ell \geq 2$ and $\{Z_t\}_{t=1}^{\ell}$ be i.i.d. random variables with distribution Z satisfying $|Z_t| \leq b$ for all $t \in [\ell]$ Let the sample variance be defined as,*

$$\mathrm{Var}_\ell(Z) = \frac{1}{\ell(\ell-1)} \sum_{1 \leq i < j \leq \ell} (Z_i - Z_j)^2$$

With probability at least $1-\delta$,

$$\mathbb{E}[Z] - \frac{1}{\ell}\sum_{i=1}^{\ell} Z_i \leq \sqrt{\frac{4\mathrm{Var}_\ell(Z)\ln\frac{4\ell^2}{\delta}}{\ell}} + \frac{7b\ln\frac{4\ell^2}{\delta}}{3(\ell-1)}$$

for all $\ell \in [\mathbb{N}]$.

Lemma C.4 (Lemma 7.5 of [2]). *Consider arbitrary T sequence of trajectories $\tau_t = \{s_h^t, a_h^t\}_{h=1}^H$, for $t = 0, 1, \ldots, T$. We have*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbf{1}((s_h^t, a_h^t) \in \mathcal{U})}{\sqrt{N_h^t(s_h^t, a_h^t)}} \leq 2\sqrt{|\mathcal{U}| \sum_{(s,a)\in\mathcal{U}} N_{h(s)}^T(s,a)} \leq 2\sqrt{|\mathcal{U}|T}. \tag{19}$$

Where $\mathcal{U} \subseteq \mathcal{S} \times \mathcal{A}$ is an arbitrary set of state action pairs and $h(s)$ corresponds to horizon index of the state partitions that contains state s.

Proof. Consider swapping the order of summation

$$\begin{aligned} \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} &= \sum_{h=1}^{H}\sum_{t=1}^{T} \frac{1}{\sqrt{N_h^t(s_h^t, a_h^t)}} = \sum_{(s,a)\in\mathcal{U}} \sum_{i=1}^{N_{h(s)}^T(s,a)} \frac{1}{\sqrt{i}} \\ &\leq 2 \sum_{(s,a)\in\mathcal{U}} \sqrt{N_{h(s)}^T(s,a)} \leq 2\sqrt{|\mathcal{U}| \sum_{(s,a)\in\mathcal{U}} N_{h(s)}^T(s,a)} \leq 2\sqrt{|\mathcal{U}|T}, \end{aligned} \tag{20}$$

where the first inequality use the fact that $\sum_{i=1}^N 1/\sqrt{i} \leq 2\sqrt{N}$ and the second inequality holds due to Cauchy-Schwarz inequality. □

Lemma C.5 (Sum of inverse counts). *Consider arbitrary T sequence of trajectories $\tau_t = \{s_h^t, a_h^t\}_{h=1}^H$, for $t = 0, 1, \ldots, T$. We have*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbf{1}((s_h^t, a_h^t) \in \mathcal{U})}{N_h^t(s_h^t, a_h^t)} \leq 2|\mathcal{U}|\log(T+1). \tag{21}$$

Where $\mathcal{U} \subseteq \mathcal{S} \times \mathcal{A}$ is an arbitrary set of state action pairs.

Proof. Consider swapping the order of summation

$$\begin{aligned} \sum_{t=1}^{T}\sum_{h=1}^{H} \frac{1}{N_h^t(s_h^t, a_h^t)} &= \sum_{h=1}^{H}\sum_{t=1}^{T} \frac{1}{N_h^t(s_h^t, a_h^t)} = \sum_{(s,a)\in\mathcal{U}} \sum_{i=1}^{N_{h(s)}^T(s,a)} \frac{1}{i} \\ &\leq 2 \sum_{(s,a)\in\mathcal{U}} \log\left(N_{h(s)}^T(s,a) + 1\right) \leq 2|\mathcal{U}|\log(T+1) \end{aligned} \tag{22}$$

□

D Supporting Algorithms

The UCBVI algorithm [8, 2] is described below.

Algorithm 4 UCBVI

```
Input reward function $r$ (assumed to be known), confidence parameters
for $t = 1, ..., T$
    Compute $\widehat{P}_t$ using all previous empirical transition data as $\widehat{\mathbb{P}}_t(s'|s,a) := \frac{N_h^t(s,a,s')}{N_h^t(s,a)}$, $\forall h, s, a, s'$.
    Compute reward bonus $b_h^t(s,a) = 2H\sqrt{\ln(SAHK/\delta)/N_h^t(s,a)}$
    Run Value-Iteration on $\{\widehat{\mathbb{P}}_t, r + b_h^t\}_{h=0}^{H-1}$.
    Set $\pi_t$ as the returned policy of VI.
End for
```

UCBVI works by adding a count based bonus to the rewards and then running value iteration over the empirical model based on these bonus augmented rewards. This encourages optimism of the empirical value functions which ensures an appropriate trade-off between exploration and exploitation. In contrast with previous approaches with provable guarantees for tabular RL problems such as UCRL [6].

As described in [8], the sample complexity of UCBVI is (up to low order terms) $\widetilde{\mathcal{O}}\left(H\sqrt{|\mathcal{S}||\mathcal{A}|}\right)$ with variance aware bonuses and $\widetilde{\mathcal{O}}\left(H^{3/2}\sqrt{|\mathcal{S}||\mathcal{A}|T}\right)$ when the bonus terms are computed using simple count based scores as above. Our results are based on a modification of the rates derived in [2] that are of order $\widetilde{\mathcal{O}}(H^2|\mathcal{S}|\sqrt{|\mathcal{A}|T})$. Despite this base rate having a suboptimal dependence in H and $|\mathcal{S}|$ in contrast with the minimax rate of [8] our results still represent an big improvement w.r.t. the un-shaped rates of [8] when the effective state space is small.

E Experimental Details

We conducted experiments building on the UCBVI framework outlined in [8]. Specifically, we used the bonus 1 formulation from the UCBVI Algorithm 4 in [8], with the bonus specifically being $\frac{C}{\sqrt{N(s,a)}}$ as the base UCBVI implementation. Our implementation is based on the open source implementation by Ian Osband `https://github.com/iosband/TabulaRL`, and open source code is attached.

The implementations of UCBVI-BS replaces this bonus by $\frac{C.\widetilde{V}}{\sqrt{N(s,a)}}$. In our implementation we chose the scaling to be 0.1 with a hyperparameter sweep. The $\widetilde{V}$ was chosen by scaling the optimal value function $V^\star$ by factors chosen between $1+\beta$ and $1-\beta$, with *beta* as $0.2, 0.5, 0.9$ (as described in the experimental section).

The environments themselves are open gridworlds of size 8, a corridor of size $10x10$ and the double corridor has size $10x20$. The reward is 0 everywhere except the goal location. Every experiment is averaged over 3 random seeds.

F Intuition for PseudoSub_Δ and $\text{PathPseudoSub}_\Delta$

Here we provide some intuition for the concepts of PseudoSub_Δ and $\text{PathPseudoSub}_\Delta$ in a deterministic chain environment (actions being left and right with deterministic transitions). The environment has a reward of 1 at state 0 and 0 elsewhere. Discount factor is 0.8 here, and the $\widetilde{V}$ has a sandwich factor $\beta = 0.3$. The true optimal value function is shown at the top of Fig. 8. The reward shaping is shown in the middle row of Fig. 8. The concept of pseudosub is shown via the arrows at (s,a) pairs that lie in PseudoSub_Δ for this particular choice of $\widetilde{V}$ and Δ. They basically denote states which are suboptimal with Δ confidence according to the shaping function $\widetilde{V}$. Now given this definition of PseudoSub_Δ, the corresponding $\text{PathPseudoSub}_\Delta$ is shown in the last row. These are states that can be eliminated via reward shaping. As we can see, around half of the states can be eliminated this way, making exploration much more directed.

G Results for Corrupted $\widetilde{V}$

We evaluate the behavior of UCBVI-Shaped as we use a corrupted version of $\widetilde{V}$ for the shaping in Fig. 9. In particular, we do this by constructing the sandwiched value function $\widetilde{V}$ as described in our experimental evaluation (in this case with $\beta = 1.5$ in the corridor environment) and then adding gaussian noise to corrupt the value function. We experiment with corruptions of 0., 0.1, 0.5, 1.0 variance, zero-centered gaussian noise. The resulting regret is seen in Fig. 9. As expected, the results show that corruption of $\widetilde{V}$ hurts the results as more and more corruption happens, but still performs better than without any shaping at all.

H Results for RND

We also ran numerical simulations (in Fig. 10) with a neural network based “pseudo-count" method - random network distillation (RND) [12]. This trains a neural network against a random pre-initialized

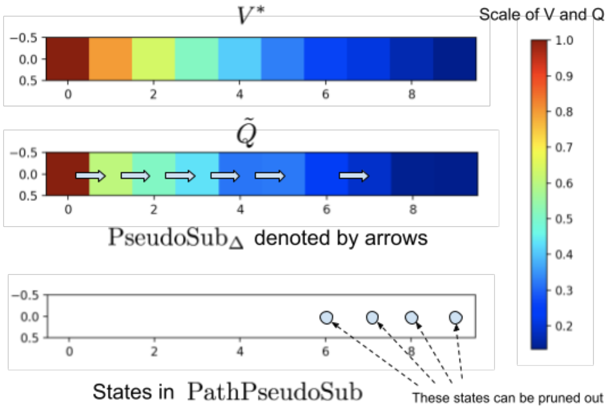

V^*
−0.5
0.0
0.5
0
2
4
6
8
$\tilde{Q}$
−0.5
0.0
0.5
0
2
4
6
8
$\mathrm{PseudoSub}_\Delta$ denoted by arrows
−0.5
0.0
0.5
0
2
4
6
8
States in $\mathrm{PathPseudoSub}$
These states can be pruned out
Scale of V and Q
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2

Figure 8: Illustration of the concepts of $\mathrm{PseudoSub}_\Delta$ and $\mathrm{PathPseudoSub}_\Delta$. Top row shows the optimal value function for this chain MDP (colorbar shows scale). Middle row shows the $\tilde{Q}$ obtained from the shaping and the (s,a) pairs which are in $\mathrm{PseudoSub}_\Delta$ (denoted by arrows). From $\mathrm{PseudoSub}_\Delta$, the bottom row shows states in $\mathrm{PathPseudoSub}_\Delta$ which can be pruned out via shaping.

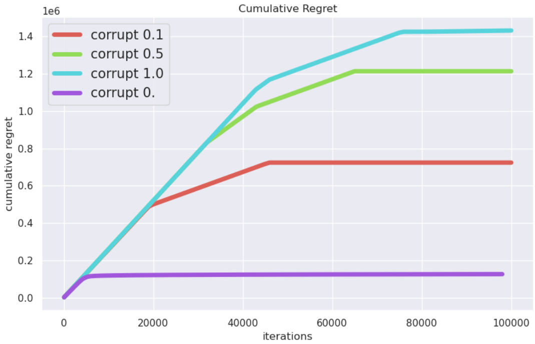


Figure 9: Illustration of behavior with corrupted shaping functions $\widetilde{V}$ with varying levels of corruption

target and uses projection error as a “pseudocount". We ran UCBVI-Shaped-BS with the inverse counts typical in UCBVI-shaped replaced with model error. The results in Fig. 10 are considerably slower than with exact counts likely due to noise in counts estimation, although the results are likely to be improved with environment specific tuning.

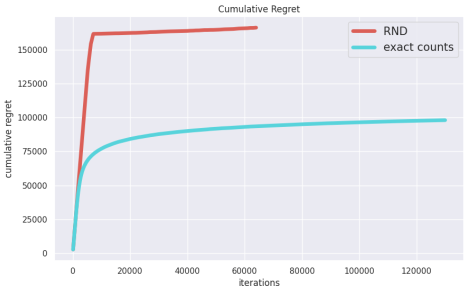


Figure 10: Illustration of behavior of UCBVI-Shaped-BS with RND based pseudocounts instead of exact counts.

I Discussion of Applicability of Reward Shaping Assumption

There can't be any way to obtain an improvement to the UCBVI regret bounds by using a reward shaping mechanism unless there is some information about the optimal policy / optimal value function encoded in the shaping function. In other words, there cannot be any free lunch in this setting. Prior knowledge is fundamental for reward shaping to be successful.

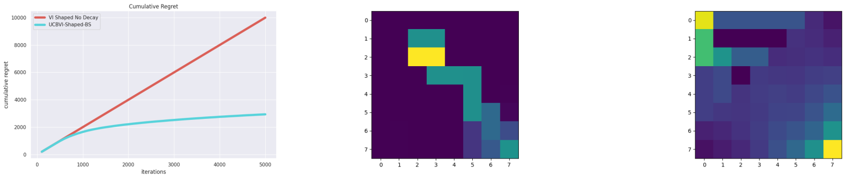
In this work we have chosen to make the assumption that the side information available to us $\widetilde{V}$ can be used to sandwich the value of $V^{\star}$. There are many examples where this assumption may hold.

1. Maze like environments. In any goal metrized goal oriented environment (for example a maze) with a stay-in-place action, if the position of the goal state is (roughly) known and the reward function equals -1 at every step from a state to a neighboring state and 0 for the stay-in-place action at the goal state, using the metric distance from the current state to the goal state may be a good estimator of the number of steps needed to reach the goal ($-V^{\star}$). The accuracy of this approximation will depend on how off this heuristic is.
2. Sim to Real. We may think of training a robot in a simulation environment and use the learned optimal value function in the simulation as our input $\widetilde{V}$. Having this information at hand should make training in real much easier than from scratch depending on how accurately the $\widetilde{V}$ estimator from the simulation tracks V^{*}. Under the assumption that dynamics are lipschitz bounded, and there is bounded state estimation error, dynamics estimation error, the resulting value function also shares the sandwich property of the true value function.
3. Object manipulation. In this setting even if the objective is to manipulate a huge combination of objects, an appropriate combination of euclidean distances of objects can serve as a good proxy for $V^{\star}$. While this may underestimate the distance to the goal especially in the presence of obstacles, in most cases a sandwich term β can be found that bounds the true value function (under the assumption that the number of obstacles is bounded). This becomes particularly pronounced as the number of objects increases and there is a combinatorially large state space, all of which cannot possibly be explored.
4. Motion planning: Recent methods have attempted to combine motion planning and RL [18]. One such instantation would involve using an anytime motion planning to generate a reward function to guide RL. An anytime motion planner may start by generating suboptimal paths, and gets better over time. If the motion planning is run only for some time so as to get a beta-sandwiched suboptimal path (and hence reward shaping as a sandwich on $V^{\star}$) to guide RL, this would then provide suboptimal shaping that can then be improved with subsequent reinforcement learning as described in our work.
5. First person RPG: For games like first person RPGs, reward shaping terms may look like a sum of euclidean distances to different objects of interest. This will underestimate the true value function will sandwich the optimal value function depending on how many obstacles are in the environment. This environment will also likely benefit from pruning off large parts of the state space.

In all of these situations, it is possible to construct a multiplicative approximation to $V^{\star}$ to be used as $\widetilde{V}$.

J Limitations

A limitation of our work is that the reward shaping has to be provided before hand rather than learned or inferred, which would be very interesting to explore in future work. Additionally, the shaping bound we have is a point-wise one rather than a correlation based bound in expectation, which can be susceptible to be outlier rewards which have a large sandwich term β. This should be investigated in future work in more detail. Additionally, it is likely that reward shaping can directly allow for horizon reduction in value computation, which should be considered in future work more directly. The behavior of these types of methods under function approximation also may be quite different, which should be studied more systematically.



(a) Quantitative comparison of effect of bonus decay

(b) Visualization of non decaying agent getting stuck

(c) Visualization of decaying agent (UCBVI-Shaped) reaching goal

Figure 11: Effect of bonus decay on the performance of UCBVI-Shaped vs standard reward shaping under shaping misspecification. As we can see, without decay the agent can stuck in arbitrarily sub-optimal points, whereas with decay the agent easily converges to an optimal solution.

J.1 Does decaying suboptimal reward shaping help over standard shaped rewards?

We ran simulations to understand whether the adaptive decay of $\widetilde{V}$ by $\frac{1}{\sqrt{N(s,a)}}$ actually provides tangible benefits, specifically when the shaping is suboptimal. We compared UCBVI-Shaped with a variant where $\widetilde{V}$ is added to the reward linearly, as done in many practical RL implementations. As seen in Fig. 11, with suboptimal shaping, UCBVI-shaped dampens the effect of the suboptimal shaping and succeed, whereas simple addition of shaping leads to the agent becoming trapped. This suggests that not only can reward shaping improve sample efficiency, but incorporating reward shaping via UCBVI-Shaped can mitigate the potential bias from suboptimal shaping.