Fast Rates for Maximum Entropy Exploration

Daniil Tiapkin [1][2]   **Denis Belomestny** [3][1]   **Daniele Calandriello** [4]   **Éric Moulines** [5][6]   **Rémi Munos** [4]
Alexey Naumov [1]   **Pierre Perrault** [7]   **Yunhao Tang** [4]   **Michal Valko** [4]   **Pierre Ménard** [8]

Abstract

We address the challenge of exploration in reinforcement learning (RL) when the agent operates in an unknown environment with sparse or no rewards. In this work, we study the maximum entropy exploration problem of two different types. The first type is *visitation entropy maximization* previously considered by Hazan et al. (2019) in the discounted setting. For this type of exploration, we propose a game-theoretic algorithm that has $\widetilde{\mathcal{O}}(H^3 S^2 A/\varepsilon^2)$ sample complexity thus improving the ε-dependence upon existing results, where S is a number of states, A is a number of actions, H is an episode length, and ε is a desired accuracy. The second type of entropy we study is the *trajectory entropy*. This objective function is closely related to the entropy-regularized MDPs, and we propose a simple algorithm that has a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/\varepsilon)$. Interestingly, it is the first theoretical result in RL literature that establishes the potential statistical advantage of regularized MDPs for exploration. Finally, we apply developed regularization techniques to reduce sample complexity of visitation entropy maximization to $\widetilde{\mathcal{O}}(H^2 S A/\varepsilon^2)$, yielding a statistical separation between maximum entropy exploration and reward-free exploration.

1. Introduction

In reinforcement learning (RL), an agent interacts with an environment aiming to maximize the sum of rewards returned by the environment. When the reward signal is very sparse or completely absent, the agent may experience long periods without any feedback. In these periods exploration is the main challenge.

This work studies the problem of efficient *exploration in the absence of rewards*. Approaches to solve this problem can be roughly cast into three main groups: The *bonus-based exploration* where the agent maximizes self-defined bonuses or intrinsic rewards collected along trajectory (Schmidhuber, 1991; Oudeyer et al., 2007; Bellemare et al., 2016). Typically these bonuses are related to the variances of error-signals from some auxiliary tasks, such as learning the transition probability distributions (Schmidhuber, 1991; Chentanez et al., 2004; Pathak et al., 2017; Savas et al., 2019), learning the optimal value function for all the possible rewards (Jin et al., 2020; Kaufmann et al., 2021; Ménard et al., 2021), learning random generated features (Burda et al., 2019). A second approach is the *goal-conditioned exploration* where the agent learns to navigate to self-assigned states (or goals). A common goal-selection rule for this class of algorithms is to select as goals the states at the frontier of the visited states (Lim & Auer, 2012; Tarbouriech et al., 2020a; Ecoffet et al., 2019). Other selection-goal rules include reaching each state a fixed number of times (Tarbouriech et al., 2021) or going to states where the estimation error for the transition probabilities is large (Tarbouriech et al., 2020b). The third approach, which has received relatively less attention thus far, is the *maximum entropy exploration* (Hazan et al., 2019; Lee et al., 2019; Mutti & Restelli, 2020; Mutti et al., 2021). This approach involves learning a policy that aims to achieve a visitation distribution over state-action pairs that is as uniform as possible. One specific application of this approach is in unsupervised pretraining, where it helps to obtain a better initial policy (Seo et al., 2021; Zhang et al., 2021; Mutti et al., 2022). To achieve this goal, the approach focuses on maximizing entropy-like functionals, which is the main focus of our study.

In this work, we focus on environments modeled by an episodic, finite, reward-free Markov Decision Process (MDP) with S states, A actions, horizon H and step-dependent transitions. We consider two types of entropy: the *visitation entropy* and the *trajectory entropy*. The visitation entropy of a policy is defined as the sum of the entropies of the visitation distributions induced by the given policy at each step. The trajectory entropy of a policy is given by the entropy of a trajectory, generated when one follows the given policy and seen as one random variable on the

---

[1]HSE University [2]Artificial Intelligence Research Institute [3]Duisburg-Essen University [4]Google DeepMind [5]École Polytechnique [6]Mohamed Bin Zayed University of AI [7]IDEMIA [8]ENS Lyon. Correspondence to: Daniil Tiapkin <dtyapkin@hse.ru>.

corresponding path space. We study maximum entropy exploration under the (ε, δ)-PAC framework, that is, we want to learn, with probability $1 - \delta$, a policy leading to ε-optimal maximum visitation or trajectory entropy and using as few as possible interactions with the environment.

Visitation entropy Hazan et al. (2019) study maximum visitation entropy[1] exploration (MVEE) in the more general framework of convex MDPs where the agent wants to maximize a convex function of the visitation distribution. The authors in Hazan et al. (2019) propose to apply the Frank-Wolfe algorithm (Frank & Wolfe, 1956) to a smoothed version of the visitation entropy. Their algorithm, `MaxEnt`[2], has a sample complexity of order[3] $\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^3 + S^3/\varepsilon^6)$, that is, it needs to sample that number of trajectories in order to find an ε-optimal policy for MVEE. Later, Cheung (2019) obtains a better rate of order $\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2 + H^3 S/\varepsilon^3)$ for the `Toc-UCRL2` algorithm. Then, building on the ideas introduced by Abernethy & Wang (2017), Zahavy et al. (2021) reinterpret the `MaxEnt` algorithm as a method to compute the equilibrium of a particular game induced by the Legendre-Fenchel transform of the smoothed entropy. Using this new point of view, they propose the `MetaEnt` algorithm[4] with a sample complexity of order $\widetilde{\mathcal{O}}(H^4 S^2 A/(\delta^2\varepsilon^2) + H^3 S/\varepsilon^3)$. In this work, building on the ideas by Grünwald & Dawid (2002), we draw a connection between MVEE and another game. In this game, a *forecaster-player tries to predict the state-action pairs visited by a sampler-player* who aims at *surprising the forecaster-player by visiting not well predicted state-action pairs*. We propose the `EntGame` algorithm that tackles MVEE by solving this prediction game. We prove that `EntGame` has a sample complexity of order $\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2 + HSA/\varepsilon)$, thus improving over the previous rate in terms of its dependence of ε, see Table 1. The key technical point leading to this improvement is that, contrary to the previous algorithms, `EntGame` does not need to estimate accurately the visitation distribution of a policy at each iteration but only needs *one trajectory generated by following this policy*. Moreover, we propose `RegEntGame`, the regularized version of `EntGame`, that achieves sample complexity of order $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$, additionally improving the previous rates in S. The main technique behind this

improvement is exploiting a strong connection between visitation entropy and regularization in MDPs. As a result, we have shown that *MVEE is statistically strictly simpler than reward-free exploration* (Jin et al., 2020).

Trajectory entropy The second problem we consider is maximum trajectory entropy exploration (MTEE). The entropy of paths of a (Markovian) stochastic process was first introduced in Ekroot & Cover (1993). Intuitively maximizing the trajectory entropy of an MDP minimizes the predictability of paths. Also there is a close connection between MTEE and regularized RL, a very popular approach in practical applications of RL.

Contrary to MVEE, the optimal policy for MTEE can easily be obtained by solving *entropy-regularized Bellman equations* with *entropy of the transition probabilities as rewards*. Leveraging this observation, one can proceed in a similar way as for the best policy identification[5] (BPI, Fiechter 1994). Precisely, we propose two algorithms. The first one, `UCBVI-Ent` is the simplest one and computes a policy by solving optimistic version of the aforementioned Bellman equations and using it to interact with the environment. The algorithm stops when an upper confidence bound on the difference between the maximum trajectory entropy and the trajectory entropy of the current policy is small enough. The second algorithm, `RL-Explore-Ent`, is an adaptation of the reward-free exploration by Jin et al. (2020) to our setting. This algorithm has two phases. In the first phase, we compute a *preliminary exploration policy* which is then used to generate independent trajectories (data). In the second phase, a nearly optimal MTEE policy is obtained by solving the empirical Bellman equations with transitions estimated from the data collected in the first phase.

Interestingly, we prove that `RL-Explore-Ent` enjoys a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/\varepsilon)$. The key technical ingredients to obtain such fast rate are exploitation of the *smoothness introduced by the regularization* and the use of the explicit form of the optimal policy.

Regularized MDPs Notably we can adapt[6] our algorithms for MTEE to best policy identification in regularized MDPs (Neu et al., 2017; Geist et al., 2019). Especially, we consider the same entropy-regularized MDPs and associated Bellman equations as in `Soft Q-learning` (Fox et al., 2016; Schulman et al., 2017; Haarnoja et al., 2017) or `SAC` (Haarnoja et al., 2018), see Remark 4.3. We show that a variation of `RL-Explore-Ent` has a sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/(\varepsilon\lambda))$ for BPI and reward-free exploration in regularized MDPs, where λ is the regularization

---

[1]Note that Hazan et al. (2019) consider a slightly different entropy; see Remark 3.3.

[2]In this work we refer to `MaxEnt` as the algorithm by Hazan et al. (2019) applied to the visitation entropy and not to the reverse entropy as initially proposed by the authors.

[3]We adapt rates from the γ-discounted setting by replacing $1/(1 - \gamma)$ with H. To take into account step-dependent transitions we multiply the first order term by H^2.

[4]We call `MetaEnt` the specialization of their general Meta-algorithm to the special case of MVEE. Note that `MaxEnt`, `Toc-UCRL2`, `MetaEnt` could be seen as variations of the same algorithm. We use different names to distinguish, at least, the associated sample complexity.

[5]Where in this problem the goal is to identify the optimal policy of a given MDP (equipped with a reward function).

[6]That is replace the entropy of the transition probability by an arbitrary reward function.

Algorithm	Setting	Sample complexity
MaxEnt (Hazan et al., 2019)		$\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^3 + S^3/\varepsilon^6)$
Toc-UCRL2 (Cheung, 2019)		$\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2 + H^3 S/\varepsilon^3)$
MetaEnt (Zahavy et al., 2021)	MVEE	$\widetilde{\mathcal{O}}(H^4 S^2 A/(\delta^2\varepsilon^2) + H^3 S/\varepsilon^3)$
EntGame (this paper)		$\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2 + HSA/\varepsilon)$
RegEntGame (this paper)		$\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2 + H^8 S^4 A/\varepsilon)$
UCBVI-Ent (this paper)	MTEE	$\widetilde{\mathcal{O}}(H^3 SA/\varepsilon^2 + H^3 S^2 A/\varepsilon)$
RL-Explore-Ent (this paper)		$\widetilde{\mathcal{O}}(H^8 S^4 A/\varepsilon)$

Table 1. Sample complexities for MVEE and MTEE. We convert rates from the γ-discounted setting by replacing $1/(1-\gamma)$ with H or from the infinite horizon setting by replacing the diameter with H. To take into account step-dependent transitions we multiply the first order term by H^2. (For MetaEnt since they do not precisely specify the cost for estimating a visitation distribution we use the same $1/\varepsilon^3$ term as for Toc-UCRL2.)

parameter. In particular, it exhibits a *statistical separation between BPI in regularized MDP and BPI in the original MDP* since in this case the optimal sample complexity is of order $\widetilde{\mathcal{O}}(H^3 SA/\varepsilon^2)$ (Ménard et al., 2021; Domingues et al., 2021a). Thus, our analysis shows that regularization is an effective way to trade-off bias for sample complexity. Additionally, we show how to use entropy regularization to obtain a theoretically faster version of EntGame algorithm.

We highlight our main contributions:

- We propose the EntGame algorithm for MVEE with a sample complexity of order $\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2)$ thus significantly improving the existing complexity bounds for MVEE.

- We introduce the new MTEE setting for exploration and provide two algorithm: the UCBVI-Ent algorithm for MTEE with a sample complexity of order $\widetilde{\mathcal{O}}(H^3 SA/\varepsilon^2)$, and RL-Explore-Ent with a sample complexity of order $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/\varepsilon)$. Up to our knowledge, this is the first time that a fast rate (in $1/\varepsilon$) is obtained thanks to regularization.

- We adapt UCBVI-Ent and RL-Explore-Ent to solve the entropy-regularized MDPs with a sample complexity of order $\widetilde{\mathcal{O}}(H^3 SA/\varepsilon^2)$ and $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/(\lambda\varepsilon))$ correspondingly, where λ is the regularization parameter.

- We combine EntGame algorithm with regularization techniques, resulting in a new algorithm RegEntGame. This algorithm improves a sample complexity of EntGame to $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$ and shows statistical separation of MVEE from reward-free exploration.

2. Setting

We consider a finite episodic *reward-free* MDP $\mathcal{M} = \big(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, s_1\big)$, where $\mathcal{S}$ is the set of states of size S, $\mathcal{A}$ is the set of actions of size A, H is the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition

from state s to state s' by performing action a in step h. And s_1 is the fixed initial state.

Policy & value functions A general policy $\pi = (\psi_h)_{h\in[H]}$ is a collection of function $\psi_h : (\mathcal{S} \times \mathcal{A} \times [0,1])^{h-1} \times \mathcal{S} \times [0,1] \to \mathcal{A}$ that maps an history $I_h = (s_1, a_1, u_1, \ldots, s_{h-1}, a_{h-1}, u_{h-1})$ where u_h are i.i.d. uniformly distributed on the unit interval, a state s_h and an auxiliary independent uniformly distributed random variable u_h to an action $a_h = \psi_h(I_h, s_h, u_h)$. A policy is Markovian if the action depends only on the previous state and the auxiliary noise $a_h = \psi_h(s_h, u_h)$. In this case the policy can be represented as $\pi = (\pi_h)_{h\in[H]}$ a collection of mappings from states to probability distributions over actions $\pi_h : \mathcal{S} \to \Delta_\mathcal{A}$ for all $h \in [H]$ where $a_h = \psi_h(s_h, u_h) \sim \pi_h(s_h)$. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $(\pi_h g)(s) \triangleq \pi_h g(s) \triangleq \mathbb{E}_{a\sim\pi_h(s)}[g(s,a)]$ denotes the composition with policy π at step h. Also, for any distribution over actions $\pi \in \Delta_\mathcal{A}$ define $\pi g(s) \triangleq \mathbb{E}_{a\sim\pi}[g(s,a)]$.

Visitation distribution The *state-action visitation distribution* of policy π at step h is denoted by d_h^π, where $d_h^\pi(s,a)$ is the probability of reaching the state-action pair (s,a) at step h after policy π.

Visitation polytopes All the admissible collection of visitation distributions belong to the following polytope

$$\mathcal{K}_p \triangleq \Big\{ d = (d_h)_{h\in[H]} : \sum_{a\in\mathcal{A}} d_1(s,a) = \mathbb{1}\{s = s_1\} \, \forall s \in \mathcal{S}$$
$$\sum_{a\in\mathcal{A}} d_{h+1}(s,a) = \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} p_h(s|s',a')d_h(s',a') \, \forall s \in \mathcal{S}, \forall h \geq 1 \Big\}.$$

We also denote by $\mathcal{K}$ the set of collections of probability distributions over state-action pairs without the constraint to be a valid visitation distribution, that is

$$\mathcal{K} \triangleq \Big\{ d = (d_h)_{h\in[H]} : d_h(s,a) \geq 0, \, \forall(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$$
$$\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_h(s,a) = 1, \, \forall h \in [H] \Big\}.$$

Trajectory distribution We denote by $\mathcal{T} \triangleq (\mathcal{S} \times \mathcal{A})^H = \big\{(s_1, a_1, \ldots, s_H, a_H) : (s_h, a_h) \in \mathcal{S} \times \mathcal{A}, \forall h \in [H]\big\}$ the set of all possible trajectories. The probability to generate the trajectory $m = (s_1, a_1, \ldots, s_H, a_H)$ with he policy π is $q^\pi(m) \triangleq \pi(a_1|s_1)\prod_{h=2}^H p_{h-1}(s_h|s_{h-1}, a_{h-1})\pi_h(a_h|s_h)$. Note that the visitation distribution at step h of policy π is a marginal of the trajectory distribution $d_h^\pi(s,a) = \mathbb{E}_{(s_1,a_1,\ldots,s_H,a_H)\sim q^\pi}[\mathbb{1}\{(s,a) = (s_h, a_h)\}]$.

Counts and empirical transition probability the number of times the state action-pair (s,a) was visited in step h in the first t episodes are $n_h^t(s,a) \triangleq$

$\sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i) = (s, a)\}$. Next, we define $n_h^t(s'|s, a) \triangleq \sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}$ the number of transitions from s to s' at step h. The empirical distribution is defined as $\widehat{p}_h^t(s'|s, a) = n_h^t(s'|s, a)/n_h^t(s, a)$ if $n_h^t(s, a) > 0$ and $\widehat{p}_h^t(s'|s, a) \triangleq 1/A$ for all $s' \in \mathcal{S}$ else.

Additional notation For $n \in \mathbb{N}_+$ we define the set $[n] \triangleq \{1, \ldots, n\}$. For $n \in \mathbb{N}_+$ we denote by Δ_n the probability simplex of dimension n. For elements $(p, q) \in \Delta_n$ the entropy of p is denoted by $\mathcal{H}(p) \triangleq -\sum_{i \in [n]} p_i \log p_i$ and the Kullback-Leibler divergence between p and q by $\mathrm{KL}(p, q) \triangleq \sum_{i \in [n]} p_i \log(p_i/q_i)$. For a number x and any two number $m < M$ define $\mathrm{clip}(x, m, M) \triangleq \max(\min(x, M), m)$.

3. Visitation Entropy

In this section we focus on maximizing the visitation entropy defined below.

Visitation entropy We define the visitation entropy of a policy π denoted by as the sum of the visitation distribution entropies at each steps

$$\mathcal{H}_{\mathrm{visit}}(d^\pi) \triangleq \sum_{h=1}^{H} \mathcal{H}(d_h^\pi) \cdot$$

We denote by $\pi^{\star, \mathrm{VE}} \in \arg\max_\pi \mathcal{H}_{\mathrm{visit}}(d^\pi)$ a policy that maximizes the visitation entropy.

Maximum visitation entropy exploration In MVEE the agent interacts with the reward-free MDP $\mathcal{M}$ as follows. At the beginning of episode t, the agent picks a policy π^t based only on the transitions collected up to episode $t-1$. Then a new reward-free trajectory is sampled following the policy π^t and observed by the agent. At the end of each episode the agent can decide to stop collecting new data, according to a random stopping time τ, the stopping rule, and outputs a (general) policy $\widehat{\pi}$ based on the observed transitions. An agent for MVEE is therefore made of a triplet $((\pi^t)_{t \in \mathbb{N}}, \tau, \widehat{\pi})$.

Definition 3.1. (PAC algorithm for MVEE) An algorithm $((\pi^t)_{t \in \mathbb{N}}, \tau, \widehat{\pi})$ is (ε, δ)-PAC for MVEE if

$$\mathbb{P}\Big(\mathcal{H}_{\mathrm{visit}}\big(d^{\pi^{\star, \mathrm{VE}}}\big) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}}) \leq \varepsilon\Big) \geq 1 - \delta.$$

Our goal is to design an algorithm that is (ε, δ)-PAC for MVEE with as sample complexity τ as small as possible.

3.1. MVEE by Solving Game

Following the general framework of Hazan et al. (2019); Zahavy et al. (2021), it is possible to solve MVEE by applying the Frank-Wolfe algorithm to a smoothed version

of the visitation entropy. Interestingly, Abernethy & Wang (2017) showed that this procedure is equivalent to computing the Nash equilibrium of a particular game induced by the Legendre-Fenchel transform of the smoothed entropy. In fact, as noted by Grünwald & Dawid (2002), there exists another game naturally linked to MVEE, stated next.

Prediction game Maximum visitation entropy is the value of the following prediction game

$$\max_{d \in \mathcal{K}_p} \mathcal{H}_{\mathrm{visit}}(d) = \max_{d \in \mathcal{K}_p} \min_{\bar{d} \in \mathcal{K}} \sum_{(h,s,a)} d_h(s, a) \log \frac{1}{\bar{d}_h(s, a)}$$

$$= \min_{\bar{d} \in \mathcal{K}} \max_{d \in \mathcal{K}_p} \sum_{(h,s,a)} d_h(s, a) \log \frac{1}{\bar{d}_h(s, a)},$$

see Lemma H.1 in Appendix H for a proof. This game can be interpreted as follows. On the one hand, the min player, or forecaster player, tries to predict which state-action pairs the max player will visit to minimize $\mathrm{KL}(d_h, \bar{d}_h)$. On the other hand, the max player, or sampler player, is rewarded for visiting state-action pairs that the forecaster player did not predict correctly.

We now describe the algorithm `EntGame` for MVEE. In this algorithm, we let a forecaster player and a sampler player compete for T episodes long. Let us first define the two players.

Forecaster-player As forecaster-player we use the Mixture-Forecaster for a logarithmic loss, see Section 9 in (Cesa-Bianchi & Lugosi, 2006). Fix a prior count n_0 and their sum $t_0 = SAn_0$. The forecaster-player predicts at episode t the distributions $\bar{d}^t \in \mathcal{K}$ with $\bar{d}_h^t(s, a) = \bar{n}_h^{t-1}(s, a)/(t + t_0)$ where the pseudo counts are $\bar{n}_h^t(s, a) = n_h^t(s, a) + n_0$ and $n_h^t(s, a)$ the counts of state-action pairs visited by the sampler-player. Note that $\bar{d}_h^t$ can be seen as the posterior mean under a Dirichlet distribution prior $\mathcal{D}\mathrm{ir}(\{n_0\}_{(s,a) \in \mathcal{S} \times \mathcal{A}})$ on $\mathcal{S} \times \mathcal{A}$.

Sampler-player As sampler-player we choose the optimistic best-response. Define the optimistic Bellman equations

$$\overline{Q}_h^t(s, a) = \log \frac{1}{\bar{d}_h^{t+1}(s, a)} + \widehat{p}_h^t \overline{V}_{h+1}^t(s, a) + b_h^t(s, a),$$

$$\overline{V}_h^t(s) = \mathrm{clip}\left(\max_{a \in \mathcal{A}} \overline{Q}_h^t(s, a), 0, \log(t/n_0 + SA)H\right),$$
(1)

where $V_{H+1}^t = 0$ and b_h^t are some Hoeffding-like bonuses defined in (6) of Appendix B. The sampler player then plays $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the optimistic Q-values, that is, $\pi_h^{t+1}(s) \in \arg\max_{\pi \in \Delta_A} \pi \overline{Q}_h^t(s)$.

Sampling rule At each episode t the policy π^t of the sampler-player is used as a sampling rule to generate a new trajectory.

Decision rule After T episodes we output a non-Markovian policy $\widehat{\pi}$ defined as the mixture of the policies $\{\pi^t\}_{t\in[T]}$, that is, to obtain a trajectory from $\widehat{\pi}$ we first sample uniformly at random $t \in [T]$ and then follow the policy π^t. Note that the visitation distribution of $\widehat{\pi}$ is exactly the average $d^{\widehat{\pi}} = (1/T)\sum_{t\in[T]} d^{\pi^t}$.

Remark that the stopping rule of `EntGame` is deterministic and equals to $\tau = T$. The complete procedure is detailed in Algorithm 1.

Algorithm 1 `EntGame`

1: **Input:** Number of episodes T, prior counts n_0.
2: **for** $t \in [T]$ **do**
3:    # Forecaster-player
4:    Update pseudo counts $\bar{n}_h^{t-1}(s,a)$ and predict $\bar{d}_h^t(s,a)$.
5:    # Sampler-player
6:    Compute π^t by optimistic planning (1) with rewards $\log\left(1/\bar{d}_h^t(s,a)\right)$.
7:    # Sampling
8:    **for** $h \in [H]$ **do**
9:       Play $a_h^t \sim \pi_h^t(s_h^t)$
10:       Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
11:    **end for**
12:    Update counts and transition estimates.
13: **end for**
14: Output $\widehat{\pi}$ the uniform mixture of $\{\pi^t\}_{t\in[T]}$.

Theorem 3.2. *Fix $\varepsilon > 0$, $\delta \in (0,1)$ and $n_0 = 1$. Then under the choice*

$$T = \widetilde{\mathcal{O}}\left(\frac{H^4 S^2 A}{\varepsilon^2} + \frac{HSA}{\varepsilon}\right)$$

the algorithm `EntGame` *is (ε, δ)-PAC. See Theorem B.7 in Appendix B for a precise bound.*

Thus the sample complexity of `EntGame` is of order $\widetilde{\mathcal{O}}(H^4 S^2 A/\varepsilon^2)$. In particular, this result significantly improves over the previous rate for MTEE, see Table 1. Note that, by using Bernstein-like bonuses (Azar et al., 2017) instead of Hoeffding-like ones for the sampler-player would give a sample complexity of order $\widetilde{\mathcal{O}}(H^3 S^2 A/\varepsilon^2)$ saving one factor H. However, in the Section 5 we present a way to use regularization techniques to achieve a sample complexity of order $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$.

Space and time complexity Since `EntGame` relies on a model-based algorithm for the sampler-player, its space complexity is of order $\mathcal{O}(HS^2 A)$. Because of the value iteration performed by the sampler-player, the time-complexity of one iteration of `EntGame` is of order $\mathcal{O}(HS^2 A)$.

Remark 3.3. Note that our definition of the visitation entropy slightly differs from the one considered by Hazan et al.

(2019). Indeed, their definition, translated to the episodic setting, is the entropy of the average of the visitation distributions which is an upper bound on the average of the entropies by concavity of the entropy

$$\mathcal{H}\left(\frac{1}{H}\sum_{h=1}^{H} d_h^{\pi}\right) \geq \frac{1}{H}\mathcal{H}_{\text{visit}}(d^{\pi}).$$

Even if both definitions make sense in the episodic setting, we think ours is slightly more appropriate in the case of step-dependent transition probabilities. Indeed, in this case we want visitation distributions to be close to the uniform distribution over state-action pairs *for all steps*. Nevertheless, `EntGame` can be adapted to optimize the visitation entropy used in Hazan et al. (2019) simply by predicting $\bar{d}_h^t(s,a) = \sum_{h'=1}^{H} \bar{n}_{h'}^{t-1}(s,a)/(H(t+t_0))$ for the forecaster-player. We conjecture that the sample complexity of this adaptation of `EntGame` for the alternative entropy is again of order $\widetilde{\mathcal{O}}(HS^2 A/\varepsilon^2)$.

Comparison with `MaxEnt` and `MetaEnt` All three algorithms, `EntGame`, `MetaEnt` (Zahavy et al., 2021), `MaxEnt` (Hazan et al., 2019) rely on the same principle of computing, implicitly or explicitly, the equilibrium of a well chosen game and deduce from it an optimal policy for MVEE. One first difference between `EntGame` and its competitors lies in the choice of the game. While `MetaEnt`, `MaxEnt` consider the game induced by the Legendre-Fenchel conjugate of a smoothed visitation entropy (Zahavy et al., 2021), `EntGame` leverages the prediction game which looks more natural for MVEE. One advantage of using this game, is that it allows to avoid the need to smooth the visitation entropy because it is done implicitly by the forecaster-agent with the pseudo-counts. More importantly, `MaxEnt` and `MetaEnt` both needs to accurately estimate at each episode the visitation distributions of the sampler-player $d_h^{\pi^t}$, leading to an extra $1/\varepsilon^3$ term in the sample complexity. Whereas `EntGame` needs one trajectory from π^t since it only involves the estimation of the averages $1/T\sum_{t=1}^{T} d_h^{\pi^t}$.

4. Trajectory Entropy

In this section we focus on another type of entropy, the trajectory entropy, that can be efficiently maximized. The entropy of paths of a (Markovian) stochastic process is introduced by Ekroot & Cover (1993). It quantifies the randomness of realizations with fixed initial and final states. Later it was extended (Savas et al., 2019) to realizations that reach a certain set of states, rather than a fixed final state. This type of entropy is also closely related to the so-called entropy rate of a stochastic process.

Trajectory entropy We define the trajectory entropy of a policy π as the entropy of a trajectory generated with the

policy π

$$\mathcal{H}_{\text{traj}}(q^\pi) \triangleq \mathcal{H}(q^\pi) = \sum_{m \in \mathcal{T}} q^\pi(m) \log \frac{1}{q^\pi(m)} \,.$$

We denote by $\pi^{\star,\text{TE}} \in \arg\max_\pi \mathcal{H}_{\text{traj}}(q^\pi)$ a policy that maximizes the trajectory entropy.

Maximum trajectory entropy exploration MTEE differs from MVEE only in the choice of entropy. In particular an algorithm $((\pi^t)_{t \in \mathbb{N}_+}, \tau, \widehat{\pi})$ for MTEE is also a combination of a time dependent policy $(\pi^t)_{t \in \mathbb{N}_+}$, a stopping rule τ, and a decision rule $\widehat{\pi}$.

Definition 4.1. (PAC algorithm for MTEE) An algorithm $((\pi^t)_{t \in \mathbb{N}}, \tau, \widehat{\pi})$ is (ε, δ)-PAC for MTEE if

$$\mathbb{P}\Big(\mathcal{H}_{\text{traj}}\big(q^{\pi^{\star,\text{TE}}}\big) - \mathcal{H}_{\text{traj}}\big(q^{\widehat{\pi}}\big) \leq \varepsilon\Big) \leq 1 - \delta \,.$$

As noted by Eysenbach & Levine (2019), MTEE can also be connected to a prediction game. In this game, the forecaster-player aims to predict the whole trajectory that the sampler-player will generate. Remark that predicting the trajectory implies to predict, in particular, the visited state-action pairs but the reverse is not true in general [7]. We could then apply the same strategy as in Section 3 to solve MTEE. Nevertheless, for trajectory entropy, there is a more direct way to proceed.

Entropy regularized Bellman equations One big difference between MVEE and MTEE is that the optimal policy can be obtained by solving regularized Bellman equations. Indeed, thanks to the chain rule for the entropy, the trajectory entropy of a policy π is $\mathcal{H}_{\text{traj}}(d^\pi) = V_1^\pi(s_1)$ and the maximum trajectory entropy is $\mathcal{H}_{\text{traj}}\big(d^{\pi^{\star,\text{TE}}}\big) = V_1^\star(s_1)$ where the value functions V^π and $V^\star$ satisfy

$$\begin{aligned}
Q_h^\pi(s,a) &= \mathcal{H}\big(p_h(s,a)\big) + p_h V_{h+1}^\pi(s,a)\,, \\
V_h^\pi(s) &= \pi_h Q_h^\pi(s) + \mathcal{H}\big(\pi_h(s)\big)\,, \\
Q_h^\star(s,a) &= \mathcal{H}\big(p_h(s,a)\big) + p_h V_{h+1}^\star(s,a)\,, \\
V_h^\star(s) &= \max_{\pi \in \Delta_A} \big\{\pi Q_h^\star(s) + \mathcal{H}(\pi)\big\}\,,
\end{aligned}$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$. In particular, the maximum trajectory entropy policy is given by $\pi_h^{\star,\text{TE}}(s) = \arg\max_{\pi \in \Delta_A}(\pi Q^\star(s) + \mathcal{H}(\pi))$. It can be computed explicitly via $\pi_h^{\star,\text{TE}}(a|s) = \exp\big(Q_h^\star(s,a) - V_h^\star(s)\big)$ as well as the optimal value function $V_h^\star(s) = \log\big(\sum_{a \in \mathcal{A}} e^{Q_h^\star(s,a)}\big)$. We refer to Appendix C for a complete derivation.

We now describe our algorithm `RL-Explore-Ent`, the description of `UCBVI-Ent` is postponed to Appendix D. The idea of the algorithm is rather simple: since we need to solve regularized Bellman equations to obtain a maximum

---

[7]Indeed d_h^π are only the marginals of q^π.

trajectory entropy policy, we can 1) find a *preliminary exploration policy* π^{mix} allowing one to construct estimates of the transition probabilities which are uniformly good when computing expectations of arbitrary bounded functions over all policies (see Lemma G.6), and 2) solve the regularized Bellman equations based on the estimated model. A similar approach is used in reward-free exploration (Jin et al., 2020; Kaufmann et al., 2021; Ménard et al., 2021), and, in particular, our algorithm is close to `RF-RL-Explore` by Jin et al. (2020). However, the key difference is that in the presence of regularization a much smaller number of transitions (trajectories) needs to be collected in order to obtain a high quality policy.

Exploration phase This phase is devoted to learn a simple (non-Markovian) preliminary exploration policy π^{mix} that could be used to construct a accurate enough estimates of transition probabilities. This policy is obtained, as in `RF-RL-Explore`, by learning for each state s and step h, a policy that reliably reaches state s at step h. This can be done by running for N_0 iterations any regret minimization algorithm, e.g. `EULER` (Zanette & Brunskill, 2019), for the sparse reward function putting reward one at state s at step h and zero otherwise. The policy π^{mix} is defined as the uniform mixture of the aforementioned policies. Then policy π^{mix} is used to collect N fresh independent trajectories from the MDP.

Planning phase For the planning phase, the agent builds a transition model

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0 \\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}, \qquad (2)$$

where $n_h(s,a)$ is the number of visits of the state-action pair (s,a) at step h for these N sampled trajectories. The final policy is a solution to the empirical regularized Bellman equations

$$\begin{aligned}
\widehat{Q}_h^\star(s,a) &= \mathcal{H}\big(\widehat{p}_h(s,a)\big) + \widehat{p}_h \widehat{V}_{h+1}^\star(s,a)\,, \\
\widehat{V}_h^\star(s) &= \max_{\pi \in \Delta_A}\Big\{\pi \widehat{Q}_h^\star(s) + \mathcal{H}(\pi)\Big\}\,, \\
\widehat{\pi}_h(s) &= \arg\max_{\pi \in \Delta_A}\Big\{\pi \widehat{Q}_h^\star(s) + \mathcal{H}(\pi)\Big\}\,.
\end{aligned} \qquad (3)$$

The complete procedure is described in Algorithm 2. We now prove that, for the well-calibrated choice of N and N_0 of order $\widetilde{\mathcal{O}}(\text{poly}(S,A,H)/\varepsilon)$, the `RL-Explore-Ent` algorithm is (ε, δ)-PAC for MTEE and provide an upper bound on its sample complexity. For the proof we refer to Appendix E.

Theorem 4.2. *The algorithm* `RL-Explore-Ent` *with parameters* $N_0 = \Omega\Big(\frac{H^7 S^3 A \cdot L^3}{\varepsilon}\Big)$ *and* $N = \Omega\Big(\frac{H^6 S^3 A L^5}{\varepsilon}\Big)$ *is* (ε, δ)-PAC *for the MTEE problem, where* $L =$

$\log(SAH/(\varepsilon\delta))$. *Its total sample complexity $SHN_0 + N$ is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8 S^4 A}{\varepsilon}\right).$$

Algorithm 2 `RL-Explore-Ent`

1: **Input:** Target precision ε, target probability δ, number of episodes for simple exploration policy N_0, number of sampled trajectories N.
2: **for** $(s', h') \in \mathcal{S} \times [H]$ **do**
3:     Form rewards $r_h(s,a) = \mathbb{1}\{s = s', h = h'\}$.
4:     Run EULER (Zanette & Brunskill, 2019) with rewards r_h over N_0 episodes and collect all policies $\Pi_{s',h'}$.
5:     Modify $\pi \in \Pi_{s',h'} : \pi_{h'}(a|s') = 1/A$ for all $a \in \mathcal{A}$.
6: **end for**
7: Construct a uniform mixture policy π^{mix} over all $\{\pi \in \Pi_{s,h} : (s,h) \in \mathcal{S} \times [H]\}$.
8: Sample N independent trajectories $(z_n)_{n \in [N]}$ following π^{mix} in the original MDP.
9: Construct from $(z_n)_{n \in [N]}$ an empirical model $\widehat{p}_h$ as in (2).
10: Output policy $\widehat{\pi}$ as a solution to (3).

Remark 4.3. (On solving regularized MDPs) Interestingly, our approach for MTEE can be adapted to solve entropy-regularized MDPs. For a reward functions $(r_h)_{h \in [H]}$ and regularization parameter $\lambda > 0$, consider the regularized Bellman equations

$$Q_{\lambda,h}^\pi(s,a) = r_h(s,a) + p_h V_{\lambda,h+1}^\pi(s,a),$$
$$V_{\lambda,h}^\pi(s) = \pi_h Q_{\lambda,h}^\pi(s) + \lambda \mathcal{H}\big(\pi_h(s)\big),$$
$$Q_{\lambda,h}^\star(s,a) = r_h(s,a) + p_h V_{\lambda,h+1}^\star(s,a),$$
$$V_{\lambda,h}^\star(s) = \max_{\pi \in \Delta_A} \pi Q_{\lambda,h}^\star(s) + \lambda \mathcal{H}(\pi),$$

where $V_{\lambda,H+1}^\pi = V_{\lambda,H+1}^\star = 0$. Note that these are the Bellman equations used by `Soft Q-learning` (Fox et al., 2016; Schulman et al., 2017; Haarnoja et al., 2017) and `SAC` (Haarnoja et al., 2018) algorithms. We are interested in the best policy identification for this regularized MDP. That is finding an algorithm that will output an ε-optimal policy $\widehat{\pi}$ such that with probability $1 - \delta$, it holds $V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\widehat{\pi}}(s_1) \leq \varepsilon$ after a minimal number τ of trajectories sampled from the MDP $\mathcal{M}^r = (\mathcal{S}, \mathcal{A}, H, (p_h)_{h \in [H]}, (r_h)_{h \in [H]}, s_1)$. By using similar exploration and planning phases as in `RL-Explore-Ent`, we get an algorithm for BPI in the entropy-regularized MDP that also enjoys the fast rate of order $\widetilde{\mathcal{O}}\big(H^8 S^4 A/(\varepsilon\lambda)\big)$. Moreover, this algorithm could be used for more general types of regularization and even in reward-free setting with the same order of the sample complexity. Refer to Appendix D-E for precise statements and proofs.

We observe that the sample complexity for solving the regularized MDP is strictly smaller[8] than the sample complexity

---

[8]For small enough ε.

---

for solving the original MDP. Indeed, one needs at least $\widetilde{\mathcal{O}}(H^3 SA/\varepsilon^2)$ trajectory (Domingues et al., 2021a) to learn a policy π which value in the (unregularized) MDP is ε-close to the optimal value. Nevertheless, regularizing the MDP introduces a bias in the value function. Precisely we have for all π, $0 \leq V_{\lambda,1}^\pi(s_1) - V_1^\pi(s_1) \leq \widetilde{\mathcal{O}}(\lambda H)$ where $V_1^\pi(s_1)$ is the value function of π at the initial state and MDP $\mathcal{M}^r$. Thus, to solve BPI in $\mathcal{M}^r$ through BPI in the regularized MDP, one needs to take $\lambda = \widetilde{\mathcal{O}}(1/(H\varepsilon))$, leading to a sample complexity of order $\widetilde{\mathcal{O}}(H^9 S^4 A/\varepsilon^2)$. In particular, our fast rate for BPI in regularized MDP does not contradict the lower bound for BPI in the original MDP. However, our analysis shows that regularization is an effective way to trade-off bias for sample complexity.

Visitation entropy vs trajectory entropy We can compare the visitation entropy and the trajectory entropy with

$$\mathcal{H}_{\mathrm{traj}}(q^\pi) \leq \underbrace{\mathrm{KL}(q^\pi, \otimes_{h=1}^H d_h^\pi) + \mathcal{H}_{\mathrm{traj}}(q^\pi)}_{\mathcal{H}_{\mathrm{visit}}(d^\pi)} \leq H\mathcal{H}_{\mathrm{traj}}(q^\pi),$$

where $\otimes_{h=1}^H d_h^\pi$ is a product measure, see Lemma H.3 in Appendix H for a proof. Note also that in general the visitation distributions of an optimal policy for maximum trajectory entropy will be less 'spread' than the one of an optimal policy for MTEE, see Section 6 for an example. In particular one can prove that the optimal policy for MTEE is the uniform policy if the transitions are deterministic, see Lemma H.4 of Appendix H.

4.1. Proof Sketch

In this section we sketch the proof of Theorem D.10.

Properties of entropy We start from analysing several properties of the entropy function. First, we notice that the well-known log-sum-exp function is a convex conjugate to the negative entropy defined only on the probability simplex (Boyd & Vandenberghe, 2004)

$$F(x) \triangleq \log\left(\sum_{a \in \mathcal{A}} e^{x_a}\right) = \max_{\pi \in \Delta_A} \langle \pi, x \rangle + \mathcal{H}(\pi)$$

and extend its action to Q-functions

$$F(Q)(s) \triangleq \max_{\pi \in \Delta_A} \pi Q(s) + \mathcal{H}(\pi).$$

This definition is useful because we can rewrite the optimal value function for MTEE in real and empirical model as follows

$$V_h^\star(s) = F(Q_h^\star)(s), \quad \widehat{V}_h^{\widehat{\pi}}(s) = F(\widehat{Q}_h^{\widehat{\pi}})(s). \quad (4)$$

Additionally, we notice that the gradient of F is equal to the soft-max policy that maximizes the expressions above

$$\pi_h^\star(s) = \nabla F(Q_h^\star)(s), \quad \widehat{\pi}_h(s) = \nabla F(\widehat{Q}_h^{\widehat{\pi}})(s),$$

and, moreover, since the negative entropy $-\mathcal{H}(\pi)$ is 1-strongly convex with respect to ℓ_1 norm, gradients of F is 1-Lipschitz with respect to ℓ_∞ norm by the properties of the convex conjugate (Kakade et al., 2009). Combining the gradient properties with the smoothness definition to $Q^\star$ and $\widehat{Q}^{\widehat{\pi}}$ we obtain

$$
\begin{aligned}
F(Q_h^\star)(s) \leq F(\widehat{Q}_h^{\widehat{\pi}})(s) &+ \widehat{\pi}_h\Big(Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\Big)(s) \\
&+ \frac{1}{2}\|Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\|_\infty^2(s),
\end{aligned}
\tag{5}
$$

where $\|Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\|_\infty(s) = \max_{a\in\mathcal{A}}|Q_h^\star(s,a) - \widehat{Q}_h^{\widehat{\pi}}(s,a)|$.

Bound on the policy error Next we apply the key inequality (5) to analyze the error between the optimal policy and policy $\widehat{\pi}$. Using (4) yields

$$
\begin{aligned}
V_h^\star(s) - V_h^{\widehat{\pi}}(s) \leq \widehat{V}_h^{\widehat{\pi}}(s) &- V_h^{\widehat{\pi}}(s) + \widehat{\pi}_h\big(Q_h^\star - \widehat{Q}_h^{\widehat{\pi}}\big)(s) \\
&+ \frac{1}{2}\max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}}(s,a) - Q_h^\star(s,a)\Big)^2.
\end{aligned}
$$

Next, by definition of $\widehat{\pi}$ we have $\widehat{V}_h^{\widehat{\pi}}(s) = \mathcal{H}(\widehat{\pi}_h(s)) + \widehat{\pi}_h\widehat{Q}_h^{\widehat{\pi}}(s)$, therefore by the regularized Bellman equations

$$
\begin{aligned}
V_h^\star(s) - V_h^{\widehat{\pi}}(s) \leq \widehat{\pi}_h p_h\Big(V_{h+1}^\star &- V_{h+1}^{\widehat{\pi}}\Big)(s) \\
&+ \frac{1}{2}\max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}}(s,a) - Q_h^\star(s,a)\Big)^2.
\end{aligned}
$$

Finally, rolling out this expression we have

$$
V_1^\star(s_1) - V_1^{\widehat{\pi}}(s_1) \leq \frac{1}{2}\mathbb{E}_{\widehat{\pi}}\left[\sum_{h=1}^{H}\max_{a\in\mathcal{A}}\Big(\widehat{Q}_h^{\widehat{\pi}} - Q_h^\star\Big)^2(s_h,a)\right].
$$

Next we may notice that in the generative model setting[9] there is available results that tells us that $\widetilde{\mathcal{O}}(1/\varepsilon)$ samples are enough to obtain $\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^\star\|_\infty \lesssim \sqrt{\varepsilon}$ (Azar et al., 2013), and we can conclude the statement. However, in the online setup the situation is more complicated, and we apply reward-free techniques developed by Jin et al. (2020) to obtain a "surrogate" of the generative model.

5. Faster Rates for Visitation Entropy

In this section, we show how to combine the regularization techniques developed in Section 4 with `EntGame` algorithm presented in Section 3.1.

The new algorithm `RegEntGame` is based on exactly the same game-theoretical framework as `EntGame`, but uses a *regularized sampler player* instead of the usual one.

Regularized sampler-player For the sampler player, we shall take advantage of strong convexity of the visitation entropy. Beforehand, we construct an estimate of the model $\{\widehat{p}_h\}_{h\in[H]}$ by reward-free exploration, using HSN_0 samples to compute a policy π^{mix} and N samples to estimate

---

[9]When there is a sampling oracle for each state-action pair.

transitions. Next, define the empirical regularized Bellman equations

$$
\widehat{Q}_h^t(s,a) = \log\left(\frac{1}{\widehat{d}_h^{t+1}(s)}\right) + \widehat{p}_h\widehat{V}_{h+1}^t(s,a),
$$

$$
\widehat{V}_h^t(s) = \max_{\pi\in\Delta_\mathcal{A}}\{\pi\widehat{Q}_h^t(s) + \mathcal{H}(\pi)\},
$$

where $\widehat{V}_{H+1}^t = 0$. The sampler player then follows the distribution $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the regularized empirical Q-values, that is, $\pi_h^{t+1}(s) \in \arg\max_{\pi\in\Delta_A}\{\pi\widehat{Q}_h^t(s) + \mathcal{H}(\pi)\}$.

Theorem 5.1. *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. For $n_0 = 1$,*

$$
N_0 = \Omega\left(\frac{H^7 S^3 A \cdot L^3}{\varepsilon}\right), \quad N = \Omega\left(\frac{H^6 S^3 A L^5}{\varepsilon}\right),
$$

and

$$
T = \Omega\left(\frac{H^3 SAL^3}{\varepsilon^2} + \frac{H^2 S^2 A^2 L^2}{\varepsilon}\right).
$$

with $L = \log(SAH/\delta\varepsilon)$ the algorithm `RegEntGame` is (ε,δ)-PAC. The total sample complexity is equal to $SH \cdot N_0 + N + T$, that is,

$$
\tau = \widetilde{\mathcal{O}}\left(\frac{H^2 SA}{\varepsilon^2} + \frac{H^8 S^4 A}{\varepsilon}\right).
$$

Thus, the sample complexity of `RegEntGame` is of order $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$ for ε large enough. In particular, this result significantly improves over the previous rates for MVEE, see Table 1. Moreover, this result shows a rate separation between reward-free exploration (Jin et al., 2020), where the established lower bound on sample complexity $\Omega(H^3 S^2 A/\varepsilon^2)$ scales with S^2, and the visitation entropy maximization problem.

Proof idea The main proof idea is to exploit not just strong convexity of the visitation entropy with respect to Euclidean distance but its strong convexity *with respect to trajectory entropy* (Bauschke et al., 2017). It allows us to use entropy regularization as described in Section 4.1 for the sampler player resulting in an averaged regret less than ε for only $\widetilde{\mathcal{O}}(\mathrm{poly}(S,A,H)/\varepsilon)$ samples. Thus, the density estimation error becomes the leading term in the full error decomposition. For more details refer to Appendix F.

6. Experiments

In this section we report experimental results on simple tabular MDP for presented algorithms and show the difference between visitation and trajectory entropies. In particular, we compare `EntGame` and `UCBVI-Ent` algorithms with (a) random agent that takes all actions uniformly at random, (b) an optimal MVEE policy computed by solving the convex program, and (c) an optimal MTEE policy computed by solving the regularized Bellman equations. As an MDP, we
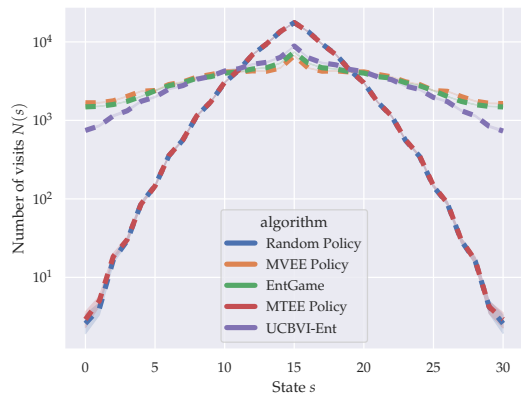
Figure 1. Number of state visits for $N = 100000$ samples in Double Chain MDP with $S = 31$ states, $A = 2$ actions, a horizon $H = 20$ and a 0.1 probability of moving to the opposite direction.

choose a stochastic environment called Double Chain as considered by Kaufmann et al. (2021).

Since the transition kernel for this environment is stage-homogeneous, for `EntGame` and `UCBVI-Ent` algorithms we joint counters over the different steps h. In particular, it changes the objective of the `EntGame` algorithm to the objective considered by Hazan et al. (2019) that makes more sense in the stage-independent setting [10].

In Figure 1 we present the number of state visits for our algorithms and baselines during $N = 100000$ interactions with the environment. For `UCBVI-Ent` algorithm the procedure was separated on two stages: at first we learn MDP with N-sample budget and extract the final policy, and then plot the number of visits for the final policy during another N samples.

In particular, we see that since this environment is almost deterministic the optimal MTEE policy is almost coincides with a random policy. Notably, the policy induced by `UCBVI-Ent` is more uniform over states because of $1/\sqrt{n^t(s,a)}$ bonuses, that make our algorithm close to `RF-UCRL` (Kaufmann et al., 2021). Also we see that the optimal MVEE policy is the most uniform over states, that makes it an appropriate target for the pure exploration problem. For more details and additional experiments we refer to Appendix I.

7. Conclusion

In this work we studied MVEE for which we provided the `EntGame` algorithm with a sampling complexity significantly smaller than the existing complexity rates. We also introduced the MTEE problem where the optimal policy can be found using the dynamic programming. We proposed the `UCBVI-Ent` and `RL-Explore-Ent` algorithms for MTEE

---

[10]See Remark 3.3.

that can be adapted to BPI in regularized MDPs. We proved that, in both cases, `RL-Explore-Ent` and its variant enjoy a fast rate. In particular, we observed a statistical separation between BPI in regularized MDP and in the original MDP. Moreover, we show that the regularized version of `EntGame` called `RegEntGame` enjoys $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$ rates, making dependence in S smaller than in the reward-free exploration problem (Jin et al., 2020).

This work opens the following interesting future research directions:

Optimal rates for MVEE and MTEE   We are still lacking lower bounds for MTEE and MVEE problems enabling us to determine the optimal rates, especially what the number of states S and the horizon H is concerned. Note that one cannot apply directly the usual lower-bounds techniques for these two problems because of the entropy regularization in both cases. In particular, we conjecture that the optimal rate for MVEE is also of order $\widetilde{\mathcal{O}}(\text{poly}(H, S, A)/\varepsilon)$.

Optimal rate for entropy-regularized RL   It would be interesting to obtain the optimal rate for BPI in a regularized MDP. In particular to recover the optimal rate for BPI in the original MDP by tuning the regularization parameter λ. We conjecture that the optimal rate is $\widetilde{\mathcal{O}}(H^2 SA/(\lambda\varepsilon))$ for BPI in entropy-regularized MDP.

Other types of entropies   Our methodology can be applied to other types of entropies and even to other regularization penalties. One interesting case would be the goal-conditioned trajectory entropy (see Savas et al. 2019) where one considers only process realizations that reach a certain set of states at time H. This entropy can be applied to goal-conditioned RL. Another type of problem that could be of high interest is visitation entropy maximization under safety constraints (Yang & Spaan, 2023).

Acknowledgements

References

Abernethy, J. D. and Wang, J.-K. On Frank-Wolfe and equilibrium computation. In *Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/file/7371364b3d72ac9a3ed8638e6f0be2c9-Paper.pdf.

Antos, A. and Kontoyiannis, I. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4): 163–193, 2001. doi: https://doi.org/10.1002/rsa.10019. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10019.

Azar, M. G., Munos, R., and Kappen, H. J. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013. URL https://hal.archives-ouvertes.fr/hal-00831875.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017. URL https://arxiv.org/pdf/1703.05449.pdf.

Bauschke, H. H., Bolte, J., and Teboulle, M. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817. URL https://doi.org/10.1287/moor.2016.0817.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/afda332245e2af431fb7b672a68b659d-Paper.pdf.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357, nov 2015. ISSN 1935-8237. doi: 10.1561/2200000050. URL https://doi.org/10.1561/2200000050.

Burda, Y., Edwards, H., Storkey, A. J., and Klimov, O. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games.* Cambridge University Press, 2006. ISBN 978-0-511-54692-1.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. Boltzmann exploration done right. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6287–6296, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Chentanez, N., Barto, A., and Singh, S. Intrinsically motivated reinforcement learning. In Saul, L., Weiss, Y., and Bottou, L. (eds.), *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004. URL https://proceedings.neurips.cc/paper/2004/file/4be5a36cbaca8ab9d2066debfe4e65c1-Paper.pdf.

Cheung, W. C. Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *CoRR*, abs/1905.06466, 2019. URL http://arxiv.org/abs/1905.06466.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2006. URL https://www.amazon.com/Elements-Information-Theory-Telecommunications-Processin dp/0471241954.

Dann, C., Lattimore, T., and Brunskill, E. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Neural Information Processing Systems*, 2017. URL https://arxiv.org/pdf/1703.07710.pdf.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, pp. 578–598. PMLR, 2021a.

Domingues, O. D., Menard, P., Pirotta, M., Kaufmann, E., and Valko, M. Kernel-based reinforcement learning: A finite-time analysis. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2783–2792. PMLR, 18–24 Jul 2021b. URL https://proceedings.mlr.press/v139/domingues21a.html.

Ecoffet, A., Huizinga, J., Lehman, J., Stanley, K. O., and Clune, J. Go-explore: a new approach for hard-exploration problems. *CoRR*, abs/1901.10995, 2019. URL http://arxiv.org/abs/1901.10995.

Ekroot, L. and Cover, T. M. The entropy of markov trajectories. *IEEE Transactions on Information Theory*, 39(4): 1418–1421, 1993.

Eysenbach, B. and Levine, S. If maxent RL is the answer, what is the question? *CoRR*, abs/1910.01913, 2019. URL http://arxiv.org/abs/1910.01913.

Fiechter, C.-N. Efficient reinforcement learning. In *Conference on Learning Theory*, 1994. URL http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=7F5F8FCD1AA7ED07356410DDD5B384FE?doi=10.1.1.49.8652&rep=rep1&type=pdf.

Fox, R., Pakman, A., and Tishby, N. Taming the noise in reinforcement learning via soft updates. In Ihler, A. and Janzing, D. (eds.), *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*. AUAI Press, 2016. URL http://auai.org/uai2016/proceedings/papers/219.pdf.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3 (1-2):95–110, 1956. doi: https://doi.org/10.1002/nav.3800030109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800030109.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized Markov decision processes. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Grill, J.-B., Darwiche Domingues, O., Menard, P., Munos, R., and Valko, M. Planning in entropy-regularized markov decision processes and games. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/50982fb2f2cfa186d335310461dfa2be-Paper.pdf.

Grünwald, P. D. and Dawid, A. P. Game theory, maximum generalized entropy, minimum discrepancy, robust bayes and pythagoras. In *Proceedings of the 2002 IEEE Information Theory Workshop, ITW 2002, 20-25 October 2002, Bangalore, India*, pp. 94–97. IEEE, 2002. doi: 10.1109/ITW.2002.1115425. URL https://doi.org/10.1109/ITW.2002.1115425.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 1352–1361. JMLR.org, 2017.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Hazan, E., Kakade, S., Singh, K., and Van Soest, A. Provably efficient maximum entropy exploration. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hazan19a.html.

Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. Reward-free exploration for reinforcement learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

Jonsson, A., Kaufmann, E., Ménard, P., Darwiche Domingues, O., Leurent, E., and Valko, M. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 33:1253–1263, 2020.

Kakade, S., Shalev-Shwartz, S., Tewari, A., et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf*, 2(1):35, 2009.

Kaufmann, E., Ménard, P., Darwiche Domingues, O., Jonsson, A., Leurent, E., and Valko, M. Adaptive reward-free exploration. In Feldman, V., Ligett, K., and Sabato, S. (eds.), *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 865–891. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/kaufmann21a.html.

Lee, L., Eysenbach, B., Parisotto, E., Xing, E. P., Levine, S., and Salakhutdinov, R. Efficient exploration via state marginal matching. *CoRR*, abs/1906.05274, 2019. URL http://arxiv.org/abs/1906.05274.

Lim, S. H. and Auer, P. Autonomous exploration for navigating in mdps. In *Conference on Learning Theory*, pp. 40–1, 2012.

Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. Fast active learning for pure exploration in reinforcement learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7599–7608. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/menard21a.html.

Mutti, M. and Restelli, M. An intrinsically-motivated approach for learning highly exploring and fast mixing policies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5232–5239, Apr. 2020. doi: 10.1609/aaai.v34i04.5968. URL https://ojs.aaai.org/index.php/AAAI/article/view/5968.

Mutti, M., Pratissoli, L., and Restelli, M. Task-agnostic exploration via policy gradient of a non-parametric state entropy estimate. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 9028–9036. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17091.

Mutti, M., Mancassola, M., and Restelli, M. Unsupervised reinforcement learning in multiple environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7850–7858, Jun. 2022. doi: 10.1609/aaai.v36i7.20754. URL https://ojs.aaai.org/index.php/AAAI/article/view/20754.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *CoRR*, abs/1705.07798, 2017. URL http://arxiv.org/abs/1705.07798.

Oudeyer, P., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.*, 11(2):265–286, 2007. doi: 10.1109/TEVC.2006.890271. URL https://doi.org/10.1109/TEVC.2006.890271.

Paninski, L. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 06 2003. ISSN 0899-7667. doi: 10.1162/089976603321780272. URL https://doi.org/10.1162/089976603321780272.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Savas, Y., Ornik, M., Cubuktepe, M., Karabag, M. O., and Topcu, U. Entropy maximization for markov decision processes under temporal logic constraints. *IEEE Transactions on Automatic Control*, 65(4):1552–1567, 2019.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991.

Schulman, J., Abbeel, P., and Chen, X. Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440, 2017. URL http://arxiv.org/abs/1704.06440.

Seo, Y., Chen, L., Shin, J., Lee, H., Abbeel, P., and Lee, K. State entropy maximization with random encoders for efficient exploration. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9443–9454. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/seo21a.html.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958. doi: pjm/1103040253. URL https://doi.org/.

Sutton, R. S. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Porter, B. and Mooney, R. (eds.), *Machine Learning Proceedings 1990*, pp. 216–224. Morgan Kaufmann, San Francisco (CA), 1990. ISBN 978-1-55860-141-3. doi: https://doi.org/10.1016/B978-1-55860-141-3.50030-4. URL https://www.sciencedirect.com/science/article/pii/B9781558601413500304.

Talebi, M. S. and Maillard, O.-A. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pp. 770–805, 2018.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. Improved sample complexity for incremental autonomous exploration in mdps. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11273–11284. Curran Associates, Inc., 2020a. URL https:

//proceedings.neurips.cc/paper/2020/file/
81e793dc8317a3dbc3534ed3f242c418-Paper.pdf.

Tarbouriech, J., Shekhar, S., Pirotta, M., Ghavamzadeh, M., and Lazaric, A. Active model estimation in markov decision processes. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1019–1028. PMLR, 03–06 Aug 2020b. URL https://proceedings.mlr.press/v124/tarbouriech20a.html.

Tarbouriech, J., Pirotta, M., Valko, M., and Lazaric, A. A provably efficient sample collection strategy for reinforcement learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7611–7624. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/3e98410c45ea98addec555019bbae8eb-Paper.pdf.

Tirinzoni, A., Al-Marjani, A., and Kaufmann, E. Optimistic pac reinforcement learning: the instance-dependent view. In *International Conference on Algorithmic Learning Theory*, pp. 1460–1480. PMLR, 2023.

van Handel, R. Probability in high dimensions. manuscript, 2014, 2014.

Yang, Q. and Spaan, M. T. Cem: Constrained entropy maximization for task-agnostic safe exploration. In *The Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023.

Zahavy, T., O'Donoghue, B., Desjardins, G., and Singh, S. Reward is enough for convex MDPs. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ELndVeVA-TR.

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7304–7312. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zanette19a.html.

Zhang, C., Cai, Y., Huang, L., and Li, J. Exploration by maximizing renyi entropy for reward-free rl framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10859–10867, May 2021. doi: 10.1609/aaai.v35i12.17297. URL https://ojs.aaai.org/index.php/AAAI/article/view/17297.

Appendix

Table of Contents

A. Notation

Table 2. Table of notation use throughout the paper

Notation	Meaning
$\mathcal{S}$	state space of size S
$\mathcal{A}$	action space of size A
H	length of one episode
s_1	initial state
τ	stopping time
$\mathcal{T}$	trajectory space, $\mathcal{T} \triangleq (\mathcal{S} \times \mathcal{A})^H$
ε	desired accuracy of solving the problem
δ	desired upper bound on failure probability
λ	regularization parameter in regularized MDPs (see Appendix D).
κ	weight of transition entropy in reward in regualrized MDPs (see Appendix D)
$p_h(s'\|s,a)$	probability transition
$r_h(s,a)$	reward function
$d_h^\pi(s,a)$	state-action visitation distribution at step h for the policy π
$q^\pi(m)$	visitation probability of trajectory $m \in \mathcal{T}$ by policy π
$\mathcal{K}_p$	polytope of all admissible state-action visitation distributions
$\mathcal{K}$	polytope of all admissible distributions over state-actions, $\mathcal{K} \triangleq (\Delta_{SA})^H$
$\mathcal{H}_{\text{visit}}(d^\pi)$	visitation entropy $\mathcal{H}_{\text{visit}}(d^\pi) \triangleq \sum_{h=1}^H \mathcal{H}(d_h^\pi)$ for $d^\pi \in \mathcal{K}_p$,
$\pi^{\star,\text{VE}}$	policy that maximizes $\mathcal{H}_{\text{visit}}(d^\pi)$, a solution to the MVEE problem
$\mathcal{H}_{\text{traj}}(q^\pi)$	trajectory entropy $\mathcal{H}_{\text{traj}}(q^\pi) \triangleq \mathcal{H}(q^\pi)$
$\pi^{\star,\text{TE}}$	policy that maximizes $\mathcal{H}_{\text{traj}}(q^\pi)$, a solution to the MTEE problem
n_0	number of prior visits for the forecaster-player in `EntGame`
t_0	total number of prior visits
s_h^t	state that was visited at h step during t episode
a_h^t	action that was picked at h step during t episode
$n_h^t(s,a)$	number of visits of state-action $n_h^t(s,a) = \sum_{k=1}^t \mathbb{1}\{(s_h^k, a_h^k) = (s,a)\}$
$n_h^t(s'\|s,a)$	number of transition to s' from state-action $n_h^t(s'\|s,a) = \sum_{k=1}^t \mathbb{1}\{(s_h^k, a_h^k, s_{h+1}^k) = (s,a,s')\}$.
$\overline{n}_h^t(s,a)$	pseudo number of visits of state-action $\overline{n}_h^t(s,a) = n_h^t(s,a) + n_0$
$\widehat{p}_h^t(s'\|s,a)$	empirical probability transition $\widehat{p}_h^t(s'\|s,a) = n_h^t(s'\|s,a)/n_h^t(s,a)$
$\overline{d}_h^t(s,a)$	predicted distribution by the forecaster-player in `EntGame` $\overline{d}_h^t(s,a) \triangleq \overline{n}_h^{t-1}(s,a)/(t+t_0)$
$\overline{Q}_h^t(s,a), \overline{V}_h^t(s,a)$	for `EntGame`: upper bound on the optimal Q/V-functions in a MDP with rewards $\log(1/\overline{d}_h^{t+1}(s,a)))$
$Q_h^\pi(s,a), V_h^\pi(s,a)$	Q- and V-functions for the MTEE problem
$Q_h^\star(s,a), V_h^\star(s,a)$	optimal Q- and V-function for the MTEE problem
$\overline{Q}_h^t(s,a), \overline{V}_h^t(s,a)$	for `UCBVI-Ent`: the upper bound on the optimal Q/V-functions for the MTEE problem
$\underline{Q}_h^t(s,a), \underline{V}_h^t(s,a)$	for `UCBVI-Ent`: the lower bound on the optimal Q/V-functions for the MTEE problem
$Q_{\lambda,h}^\pi(s,a), V_{\lambda,h}^\pi(s,a)$	Q- and V-functions in a regularized MDP
$Q_{\lambda,h}^\star(s,a), V_{\lambda,h}^\star(s,a)$	optimal Q- and V-function in a regularized MDP
$\widehat{Q}_{\lambda,h}^\pi(s,a), \widehat{V}_{\lambda,h}^\pi(s,a)$	for `RL-Explore-Ent`: the empirical Q- and V-functions in a regularized MDP
$\mathbb{R}_+$	non-negative real numbers
$\mathbb{N}_+$	positive natural numbers
$[n]$	set $\{1, 2, \ldots, n\}$
e	Euler's number
Δ_d	$d+1$-dimensional probability simplex: $\Delta_d \triangleq \{x \in \mathbb{R}_+^d : \sum_{j=1}^d x_j = 1\}$
$\Delta_\mathcal{X}$	set of distributions over a finite set $\mathcal{X} : \Delta_\mathcal{X} = \Delta_{\|\mathcal{X}\|}$.
$\mathcal{H}(p)$	Shannon entropy for $p \in \Delta_\mathcal{X}, \mathcal{H}(p) \triangleq \sum_{i \in \mathcal{X}} p_i \log(1/p_i)$
$\text{clip}(x, m, M)$	clipping procedure $\text{clip}(x,m,M) \triangleq \max(\min(x,M), m)$

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $\mathcal{P}(\mathsf{X})$ be the set of all probability measures on this space. For $p \in \mathcal{P}(\mathsf{X})$ we denote

by $\mathbb{E}_p$ the expectation w.r.t. p. For random variable $\xi : \mathsf{X} \to \mathbb{R}$ notation $\xi \sim p$ means $\mathrm{Law}(\xi) = p$. For any measures $p, q \in \mathcal{P}(\mathsf{X})$ we denote their product measure by $p \otimes q$. We also write $\mathbb{E}_{\xi \sim p}$ instead of $\mathbb{E}_p$. For any $p, q \in \mathcal{P}(\mathsf{X})$ the Kullback-Leibler divergence $\mathrm{KL}(p, q)$ is given by

$$\mathrm{KL}(p, q) = \begin{cases} \mathbb{E}_p[\log \frac{\mathrm{d}p}{\mathrm{d}q}], & p \ll q \\ +\infty, & \text{otherwise} \end{cases}$$

For any $p \in \mathcal{P}(\mathsf{X})$ and $f : \mathsf{X} \to \mathbb{R}$, $pf = \mathbb{E}_p[f]$. In particular, for any $p \in \Delta_d$ and $f : \{0, \ldots, d\} \to \mathbb{R}$, $pf = \sum_{\ell=0}^{d} f(\ell) p(\ell)$. Define $\mathrm{Var}_p(f) = \mathbb{E}_{s' \sim p}[(f(s') - pf)^2] = p[f^2] - (pf)^2$. For any $(s, a) \in \mathcal{S}$, transition kernel $p(s, a) \in \mathcal{P}(\mathcal{S})$ and $f \colon \mathcal{S} \to \mathbb{R}$ define $pf(s, a) = \mathbb{E}_{p(s,a)}[f]$ and $\mathrm{Var}_p[f](s, a) = \mathrm{Var}_{p(s,a)}[f]$. For any $s \in \mathcal{S}$, policy $\pi(s) \in \mathcal{P}(\mathcal{S})$ and $f \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ define $\pi f(s) = \mathbb{E}_{a \sim \pi(s)}[f(s, a)]$ and $\mathrm{Var}_\pi f(s) = \mathrm{Var}_{a \sim \pi(s)}[f(s, a)]$.

For a MDP $\mathcal{M}$, a policy π and a sequence of function f_h define $\mathbb{E}_\pi[\sum_{h'=h}^{H} f(s_{h'}, a_{h'}) | s_h]$ as a conditional expectation of $\sum_{h'=h}^{H} f(s_{h'}, a_{h'})$ with respect to the sigma-algebra $\mathcal{F}_h = \sigma\{(s_{h'}, a_{h'}) | h' \leq h\}$, where for any $h \in [H]$ we have $a_h \sim \pi(s_h), s_{h+1} \sim p_h(s_h, a_h)$.

We write $f(S, A, H, \varepsilon) = \mathcal{O}(g(S, A, H, \varepsilon, \delta))$ if there exist $S_0, A_0, H_0, \varepsilon_0, \delta_0$ and constant $C_{f,g}$ such that for any $S \geq S_0, A \geq A_0, H \geq H_0, \varepsilon < \varepsilon_0, \delta < \delta_0, f(S, A, H, T, \delta) \leq C_{f,g} \cdot g(S, A, H, T, \delta)$. We write $f(S, A, H, \varepsilon, \delta) = \widetilde{\mathcal{O}}(g(S, A, H, \varepsilon, \delta))$ if $C_{f,g}$ in the previous definition is poly-logarithmic in $S, A, H, 1/\varepsilon, 1/\delta$.

B. Proofs for Visitation Entropy

We first define the regrets of each players obtained by playing T times the games. For the forecaster-player, for any $\bar{d} \in \mathcal{K}$ we define

$$\mathfrak{R}_{\mathrm{Fore}}^T(\bar{d}) \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \widetilde{d}_h^t(s,a) \left(\log \frac{1}{\bar{d}_h^t(s,a)} - \log \frac{1}{\bar{d}_h(s,a)} \right)$$

where $\widetilde{d}_h^t(s,a) \triangleq \mathbb{1}\{(s_h^t, a_h^t) = (s,a)\}$ is a sample from $d_h^{\pi^t}(s,a)$. Similarly for the sampler-player, for any $d \in \mathcal{K}_p$ we define

$$\mathfrak{R}_{\mathrm{Samp}}^T(d) \triangleq \sum_{t=1}^{T} \sum_{h,s,a} \left(d_h(s,a) - d_h^{\pi^t}(s,a) \right) \log \frac{1}{\bar{d}_h^t(s,a)} \, .$$

Recall that the visitation distribution of the policy π returned by `EntGame` is the average of the visitation distributions of the sampler-player $d_h^{\widehat{\pi}}(s,a) = \hat{d}_h^T(s,a) \triangleq (1/T) \sum_{t=1}^{T} d_h^{\pi^t}(s,a)$. We also denote by $\mathring{d}_h^T(s,a) \triangleq (1/T) \sum_{t=1}^{T} \widetilde{d}^t(s,a)$ the average of the 'sample' visitation distributions.

We now relate the difference between the optimal visitation entropy and the visitation entropy of the outputted policy $\widehat{\pi}$ with the regrets of the two players. Indeed, using $\mathcal{H}(p) = \sum_{i \in [n]} p_i \log(1/q_i) - \mathrm{KL}(p,q) \le \sum_{i \in [n]} p_i \log(1/q_i)$ for all $(p,q) \in (\Delta_n)^2$, we obtain

$$T\left(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}}) \right) \le \sum_{t=1}^{T} \sum_{h,a,s} d_h^{\pi^\star}(s,a) \log \frac{1}{\bar{d}_h^t(s,a)} - \widetilde{d}_h^t(s,a) \log \frac{1}{\mathring{d}_h^T(s,a)} + T\left(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\hat{d}^T) \right)$$

$$= \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star}) + \underbrace{\sum_{t=1}^{T} \sum_{h,s,a} \left(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a) \right) \log \frac{1}{\bar{d}_h^t(s,a)}}_{\mathrm{Bias}_1} + \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T)$$

$$+ \underbrace{T\left(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\hat{d}^T) \right)}_{\mathrm{Bias}_2} \, .$$

It remains to upper bound each terms separately in order to obtain a bound on the gap. We first bound the two regrets terms. The first bias term is martingale and can easily be bounded with a deviation inequality, whereas for the second one we introduce just instrumentally smoothing of the entropy.

B.1. Regret of the Forecaster-Player

We prove in this section a regret-bound for the mixture forecaster.

Lemma B.1. *For $n_0 = 1$, for any $\bar{d} \in \mathcal{K}$ it holds almost surely*

$$\mathfrak{R}_{\mathrm{Fore}}^T(\bar{d}) \le HSA \log\left(\mathrm{e}(T+1) \right) - T \sum_{h=1}^{H} \mathrm{KL}(\mathring{d}_h^T, \bar{d}_h) \, .$$

Proof. We will bound the regret at step h,

$$\mathfrak{R}_{\mathrm{Fore},h}^T(\bar{d}) \triangleq \sum_{t=1}^{T} \sum_{s,a} \widetilde{d}_h^t(s,a) \left(\log \frac{1}{\bar{d}_h^t(s,a)} - \log \frac{1}{\bar{d}_h(s,a)} \right)$$

and then sum the upper bounds. Recall

$$\bar{d}_h^t(s,a) = \frac{n_h^{t-1}(s,a) + 1}{t - 1 + SA},$$

and for $(s,a) = (s_h^t, a_h^t)$ and any $t \in [T], h \in [H]$ we have $n_h^{t-1}(s,a) + 1 = n_h^t(s,a)$. Since $n_0 = 1$, we have

$\bar{d}_h^t(s_h^t, a_h^t) = n_h^t(s_h^t, a_h^t)/(SA + t - 1)$. Armed with this observation we can rewrite the regret as follows

$$\mathfrak{R}_{\text{Fore},h}^T(\bar{d}) = -T \, \text{KL}(\mathring{d}_h^T, \bar{d}_h) - T\mathcal{H}(\mathring{d}_h^T) - \sum_{t=1}^T \log\left(\bar{d}_h^t(s_h^t, a_h^t)\right)$$

$$= -T \, \text{KL}(\mathring{d}_h^T, \bar{d}_h) - T\mathcal{H}(\mathring{d}_h^T) - \log\left(\prod_{t=1}^T \bar{d}_h^t(s_h^t, a_h^t)\right).$$

Then we have an explicit formula for the product of $\bar{d}_h^t$

$$\prod_{t=1}^T \bar{d}_h^t(s_h^t, a_h^t) = \prod_{t=1}^T \frac{n_h^t(s_h^t, a_h^t)}{SA + t - 1} = \frac{(SA - 1)!}{(SA + T - 1)!} \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} [n_h^T(s, a)]!$$

$$= \frac{1}{\binom{T}{(n_h^T(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}}}} \frac{1}{\binom{T+SA-1}{SA-1}}$$

$$\geq \exp\left(-T\mathcal{H}(\mathring{d}_h^T) - (T + SA - 1)\mathcal{H}\left(\frac{SA - 1}{T + SA - 1}\right)\right)$$

where in the last inequality we used Theorem 11.1.3 by Cover & Thomas (2006) and overload the entropy notation $\mathcal{H}(p) = -p\log(p) - (1 - p)\log(1 - p)$ for $p \in [0, 1]$. Putting all together we get

$$\mathfrak{R}_{\text{Fore},h}^T(\bar{d}) \leq (T + SA - 1)\mathcal{H}\left(\frac{A - 1}{T + A - 1}\right) - T \, \text{KL}(\mathring{d}_h^T, \bar{d}_h).$$

Bounding the entropic term

$$(T + SA - 1)\mathcal{H}\left(\frac{SA - 1}{T + SA - 1}\right) = (SA - 1)\log\frac{T + SA - 1}{SA - 1} + T\log\frac{T + SA - 1}{T}$$

$$\leq (SA - 1)\log\frac{T + SA - 1}{SA - 1} + T\log\left(1 + \frac{SA - 1}{T}\right)$$

$$\leq (SA - 1)\log\frac{\text{e}(T + SA - 1)}{SA - 1}$$

$$\leq SA\log\left(\text{e}(T + 1)\right),$$

and summing over h allows us to conclude. $\qquad\square$

B.2. Regret of the Sampler-Player

We start from introducing new notation. Let $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h^t\}_{h\in[H]}, s_1)$ be a sequence of MDPs where reward defined as follows $r_h^t(s, a) = \log(1/\bar{d}_h^t(s, a))$. Define $Q_h^{\pi,t}(s, a)$ and $V_h^{\pi,t}(s, a)$ as a action-value and value functions of a policy π on a MDP $\mathcal{M}_t$. Notice that the value-function of initial state in this case could be written as follows

$$V_1^{\pi,t}(s_1) = \sum_{h,s,a} d_h^\pi(s, a)\log\left(\frac{1}{\bar{d}_h^t(s, a)}\right)$$

therefore, the regret for the sampler-player could be rewritten in the terms of the regret for this sequence of MDPs

$$\mathfrak{R}_{\text{Samp}}^T(d^\pi) = \sum_{t=1}^T V_1^{\pi,t}(s_1) - V_1^{\pi^t,t}(s_1).$$

Since the rewards are changing in each episode and depending on the full history on interaction during previous episodes, we have to handle more uniform approach as in usual UCBVI proofs (Azar et al., 2017).

Concentration   Let $\alpha^{\mathrm{KL}}, \alpha^{\mathrm{cnt}} : (0,1) \times \mathbb{R}_+ \to \mathbb{R}_+$ be some functions defined later on in Lemma B.2. We define the following favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leq \frac{\alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \right\},$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \quad n_h^t(s,a) \geq \frac{1}{2} \overline{n}_h^t(s,a) - \alpha^{\mathrm{cnt}}(\delta) \right\},$$

Lemma B.2. *For any $\delta \in (0,1)$ and for the following choices of functions α,*

$$\alpha^{\mathrm{KL}}(\delta, n) \triangleq \log(2SAH/\delta) + S \log(\mathrm{e}(1+n)), \quad \alpha^{\mathrm{cnt}}(\delta) \triangleq \log(2SAH/\delta),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1 - \delta/2, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geq 1 - \delta/2$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \geq 1 - \delta$.

Proof. Applying Theorem G.1 and the union bound over $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1 - \delta/2$. By Theorem G.3 and union bound, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geq 1 - \delta/2$. The union bound yields $\mathbb{P}[\mathcal{G}(\delta)] \geq 1 - \delta$. $\qquad\square$

Optimism   Next we define the exploration bonuses $b_h^t(s,a)$ for the sampler-player for $n_0 = 1$

$$b_h^t(s,a) = \sqrt{\frac{2H^2 \log^2(t + SA) \cdot \alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \tag{6}$$

Lemma B.3. *For any $t \in [T]$ and any policy π, the following holds on event $\mathcal{G}(\delta)$*

$$\overline{Q}_h^t(s,a) \geq Q_h^{\pi,t+1}(s,a), \qquad \overline{V}_h^t(s) \geq V_h^{\pi,t+1}(s).$$

Proof. Proceed by backward induction over h. For $h = H + 1$ the statement trivially holds. Next we assume that the statement holds for any $h' > h$. Then we have by induction hypothesis and Hölder's inequality

$$\begin{aligned}
\overline{Q}_h^t(s,a) - Q_h^{\pi,t+1}(s,a) &= \widehat{p}_h^t \overline{V}_{h+1}^t(s,a) - p_h V_h^{\pi,t+1}(s,a) + b_h^t(s,a) \\
&\geq [\widehat{p}_h^t - p_h] V_h^{\pi,t+1}(s,a) + b_h^t(s,a) \geq -\|V_h^{\pi,t+1}\|_\infty \|\widehat{p}_h^t - p_h\|_1 + b_h^t(s,a).
\end{aligned}$$

The fact that $\|V_h^{\pi,t+1}\|_\infty \leq H \log(t + SA)$, Pinsker's inequality and the definition of the event $\mathcal{E}^{\mathrm{KL}}(\delta)$ yields

$$\|V_h^{\pi,t+1}\|_\infty \|\widehat{p}_h^t - p_h\|_1 \leq H \log(t + SA) \sqrt{\frac{2\alpha^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} = b_h^t(s,a)$$

that shows $\overline{Q}_h^t(s,a) - Q_h^{\pi,t+1}(s,a) \geq 0$. The inequality on V-functions could be derived as follows

$$\overline{V}_h^t(s) \geq \pi \overline{Q}_h^t(s) \geq \pi Q_h^{\pi,t+1}(s) = V_h^{\pi,t+1}(s).$$

$\qquad\square$

Regret Bound

Lemma B.4. *Let π be any fixed policy. Then for any $\delta \in (0,1)$ with probability at least $1 - \delta$ the following holds*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leq 10 \log(T + SA) \sqrt{2H^4 SAT \cdot (\log(2SAH/\delta) + S \log(\mathrm{e}T)) \log(T)}.$$

Proof. Assume that the event $\mathcal{G}(\delta)$ holds. By Lemma B.3 for any $t \in [T]$ and $h \in [H]$ we have

$$V_t^{\pi,t}(s_h) - V_h^{\pi^t,t}(s_h^t) \leq \overline{V}_h^{t-1}(s_h^t) - V_h^{\pi^t,t}(s_h^t) = \pi_h^t(\overline{Q}_h^{t-1} - Q_h^{\pi^t,t})(s),$$

thus we can define $\delta_h^t(s,a) = \overline{Q}_h^{t-1}(s,a) - Q_h^{\pi^t,t}(s,a)$ and upper bound the regret as follows

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leq \sum_{t=1}^T \pi_1^t \delta_1^t(s_1).$$

Next we analyze $\delta_h^t(s_h^t)$. By the same argument as in Lemma B.3

$$\delta_h^t(s,a) = [\widehat{p}_h^{t-1} - p_h]\overline{V}_{h+1}^{t-1}(s,a) + b_h^t(s,a) + p_h[\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t,t}](s,a) \leq 2b_h^{t-1}(s,a) + p_h\pi_{h+1}^t[\overline{Q}_{h+1}^{t-1} - Q_{h+1}^{\pi^t,t}](s,a)$$

that could be rewritten as follows

$$\delta_h^t(s,a) \leq \mathbb{E}_{\pi^t}\left[2b_h^{t-1}(s,a) + \delta_{h+1}^t(s_{h+1},a_{h+1})|(s_h,a_h) = (s,a)\right],$$

thus, rolling out the initial bound on regret we have

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leq H\log(T + SA)\sum_{t=1}^{T-1}\mathbb{E}_{\pi^t}\left[\sum_{h=1}^H 2\sqrt{\frac{2\alpha^{\mathrm{KL}}(\delta, n_h^t(s_h,a_h))}{n_h^t(s_h,a_h)}} \wedge 1 \Bigg| s_1\right] + H\log(T + SA).$$

By Lemma H.5 and Jensen's inequality we have

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leq 5H^{3/2}\log(T + SA)\sqrt{2T}\sqrt{\sum_{h,s,a}\sum_{t=1}^{T-1} d_h^{\pi^t}(s,a)\frac{\alpha^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a) \vee 1}}.$$

Notice that $d_h^{\pi^t}(s,a) = \overline{n}_h^{t+1}(s,a) - \overline{n}_h^t(s,a)$ and $\alpha^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a)) \leq \alpha^{\mathrm{KL}}(\delta, T - 1)$. Combined with Lemma H.6 it implies

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) \leq 10\log(T + SA)\sqrt{2H^4SAT \cdot (\log(2SAH/\delta) + S\log(eT))\log(T)}.$$

Finally, the fact that $\mathbb{P}[\mathcal{G}(\delta)] \geq 1 - \delta$ concludes the statement of theorem. $\square$

Remark B.5. It is possible to improve the H-dependence by introducing Bernstein-type bonuses, however, we are focused on improvement in a dependence in ε^{-1} and leave this regret bound as simple as possible.

B.3. Bias Terms

Lemma B.6. *Let $\delta \in (0,1)$ and $n_0 = 1$. Then with probability at least $1 - \delta$ the following two bounds hold*

$$\mathrm{Bias}_1 \triangleq \sum_{t=1}^T \sum_{h,s,a}\left(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a)\right)\log\frac{1}{\overline{d}_h^t(s,a)} \leq \log(T + SA)\sqrt{2TH\log(2/\delta)}$$

$$\mathrm{Bias}_2 \triangleq T(\mathcal{H}_{visit}(\mathring{d}^T) - \mathcal{H}_{visit}(\hat{d}^T)) \leq \log(SAT)\left(\sqrt{2TH\log(2/\delta)} + 3H\sqrt{SAT\log(3T)}\right).$$

Proof. Let us define the lexicographic order on the set $[T] \times [H]$ with an additional convention $(t,0) = (t-1, H)$.

Then we can define a filtration $\mathcal{F}_{t,h} = \sigma\left\{(s_{h'}^{t'}, a_{h'}^{t'}) \forall t \leq t, \forall h' \in [H]\} \cup \{(s_{h'}^t, a_{h'}^t) \forall h' \leq h\}\right\}$ that consists of the all history of interactions of the `EntGame` algorithm with an environment up to the h-th step of the episode t. The most important fact is that π^t and $\overline{d}_h^t(s,a)$ are $\mathcal{F}_{t,h-1}$-measurable for $h > 1$ and $\mathcal{F}_{t-1,H}$-measurable for $h = 1$.

Therefore, for any $t \in [T], h \in [H]$

$$\mathbb{E}\left[\sum_{s,a}(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a))\log\frac{1}{\overline{d}_h^t(s,a)}\bigg|\mathcal{F}_{t,h-1}\right] = 0.$$

Therefore $X_{t,h} = \sum_{s,a}(d_h^{\pi^t}(s,a) - \widetilde{d}_h^t(s,a))\log\frac{1}{\widetilde{d}_h^t(s,a)}$ is a martingale-difference sequence adapted to the filtration $\mathcal{F}_{t,h}$. Also we notice that a.s. the following bound holds

$$|X_{t,h}| \le \log(T + SA)$$

All these facts combined with Azuma-Hoeffding inequality implies that with probability at least $1 - \delta/2$

$$\mathrm{Bias}_1 = \sum_{t=1}^{T}\sum_{h=1}^{H}X_{t,h} \le \log(T + SA)\sqrt{2TH\log(2/\delta)}.$$

To show the second part of the statement we notice that

$$\mathrm{Bias}_2 = T\sum_{h=1}^{H}(\mathcal{H}(\mathring{d}_h^T) - \mathcal{H}(\hat{d}_h^T)).$$

Let us introduce the smoothed entropy as it was done by Hazan et al. (2019).

$$\forall d \in \Delta_{SA} : \mathcal{H}_\sigma(d) = \sum_{s,a}d(s,a)\log\frac{1}{d(s,a) + \sigma}.$$

The key difference with our approach and approach of Hazan et al. (2019) that we need the smoothing only instrumentally to provide a bound on Bias_2.

It is easy to see that $\mathcal{H}_\sigma$ is concave and, moreover $d \in \Delta_{SA}$

$$|\mathcal{H}(d) - \mathcal{H}_\sigma(d)| \le \sum_{s,a}d(s,a)\log\frac{d(s,a) + \sigma}{d(s,a)} \le \sigma SA,$$

where we used inequality $\log(1 + x) \le x$ for all $x \ge 0$, and also for $\sigma < \mathrm{e}^{-1}$

$$\|\nabla\mathcal{H}_\sigma(d)\|_\infty = \sup_{x \in (0,1)}\left|\log(x + \sigma) + \frac{x}{x + \sigma}\right| \le \log(\sigma^{-1}).$$

By replacing an entropy with a smoothed entropy

$$\mathrm{Bias}_2 \le T\sum_{h \in H}(\mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\hat{d}_h^T)) + 2\sigma \cdot TSAH.$$

To analyze the first term we use that $\mathcal{H}_\sigma$ is concave, therefore

$$\sum_{h=1}^{H}\mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\hat{d}_h^T) \le \sum_{h=1}^{H}\langle\nabla\mathcal{H}_\sigma(\hat{d}_h^T), \mathring{d}_h^T - \hat{d}_h^T\rangle = \frac{1}{T}\sum_{s,a}\sum_{t=1}^{T}\sum_{h=1}^{H}(\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a))\cdot\nabla\mathcal{H}_\sigma(\hat{d}_h^T)(s,a)$$

For this term situation is more involved than for Bias_1 because $\hat{d}_h^T$ is dependent on all generated policies. Therefore we have to preform uniform bounds. Define $\mathcal{W} = \{w \in \mathbb{R}^{HSA} \mid |w_h(s,a)| \le 1\}$ as a unit ℓ_∞-ball in $\mathbb{R}^{HSA}$. Then we have

$$T\sum_{h=1}^{H}\mathcal{H}_\sigma(\mathring{d}_h^T) - \mathcal{H}_\sigma(\hat{d}_h^T) \le \log(\sigma^{-1})\cdot\sup_{w \in \mathcal{W}}\sum_{t=1}^{T}\sum_{h=1}^{H}\left(\sum_{s,a}(\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a))\cdot w_h(s,a)\right).$$

Define $N(\varepsilon, \|\cdot\|_\infty, \mathcal{W})$ as ε-covering number for a set $\mathcal{W}$ with ℓ_∞-norm as a distance, and $\mathcal{W}_\varepsilon$ as a minimal ε-net. Next we can use the well-known result on upper bound on the covering number (e.g. see Exercise 5.5 by van Handel (2014))

$$N(\varepsilon, \|\cdot\|_\infty, \mathcal{W}) \le \left(\frac{3}{\varepsilon}\right)^{SAH},$$

and replace our maximization problem with the maximization over ε-net

$$\sup_{w \in \mathcal{W}} \sum_{t=1}^{T} \sum_{h=1}^{H} \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot w_h(s,a) \right) \leq \sup_{\hat{w} \in \mathcal{W}_\varepsilon} \sum_{t=1}^{T} \sum_{h=1}^{H} \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot \hat{w}_h(s,a) \right) + \varepsilon T H.$$

For any fixed $\hat{w} \in \mathcal{W}_\varepsilon$ we apply Azuma-Hoeffding inequality exactly in the same manner as in the bound for Bias_1-term. We have that with probability at least $1 - \delta/(2N_\varepsilon)$ for $N = N(\varepsilon, \|\cdot\|_\infty, \mathcal{W})$ we have

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot \hat{w}_h(s,a) \right) \leq \sqrt{2TH \log(2N_\varepsilon/\delta)}.$$

Thus, by union bound we have with probability at least $1 - \delta/2$

$$\sup_{w \in \mathcal{W}} \sum_{t=1}^{T} \sum_{h=1}^{H} \left(\sum_{s,a} (\widetilde{d}_h^t(s,a) - d_h^{\pi^t}(s,a)) \cdot w_h(s,a) \right) \leq \sqrt{2TH(\log(2/\delta) + SAH \log(3/\varepsilon))} + \varepsilon T H.$$

Taking $\varepsilon = 1/T$ we have

$$\mathrm{Bias}_2 \leq \log(\sigma^{-1})(\sqrt{2TH(\log(2/\delta) + SAH \log(3T))} + H) + \sigma SATH.$$

Next we choose $\sigma = 1/SAT$ and by inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ obtain

$$\mathrm{Bias}_2 \leq \log(SAT)\left(\sqrt{2TH \log(2/\delta)} + 3H\sqrt{SAT \log(3T)} \right).$$

$\square$

B.4. Proof of Theorem 3.2

We state the version of this theorem with all prescribed dependencies factors.

Theorem B.7. *For all $\varepsilon > 0$ and $\delta \in (0,1)$. For $n_0 = 1$ and*

$$T \geq 1 + \frac{648(\log(SA) + L)H^4 SA \cdot (\log(4SAH/\delta) + S + L) \cdot L}{\varepsilon^2} + \frac{2HSA(2+L)}{\varepsilon}$$

for $L = 9\log\left(1010\sqrt{H^4 S^{8/3} A^{8/3} \log(4SAH/\delta)}/\varepsilon \right)$ the algorithm `EntGame` *is (ε, δ)-PAC.*

Proof. We start from writing down the decomposition defined in the beginning of the appendix

$$T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})) \leq \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^{\star,\mathrm{VE}}}) + \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) + \mathrm{Bias}_1 + \mathrm{Bias}_2.$$

By Lemma B.4 with probability at least $1 - \delta/2$ it holds

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^{\star,\mathrm{VE}}}) = 10\log(T+SA)\sqrt{2H^4 SAT \cdot (\log(4SAH/\delta) + S\log(eT)) \log(T)}$$

By Lemma B.1

$$\mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) \leq HSA \log\left(e(T+1) \right).$$

By Lemma B.6 with probability at least $1 - \delta/2$

$$\mathrm{Bias}_1 + \mathrm{Bias}_2 \leq 3\log(SAT)\left(\sqrt{TH \log(4/\delta)} + H\sqrt{SAT \log(3T)} \right).$$

By union bound all these inequalities hold simultaneously with probability at least $1 - \delta$. Combining all these bounds we get

$$T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})) \leq 18\log(SAT)\sqrt{H^4 SAT(\log(4SAH/\delta) + S\log(eT)) \log(T)} + HSA\log(e(T+1)).$$

Therefore, it is enough to choose T such that $\mathcal{H}_{\text{visit}}(d^{\pi^{\star,\text{VE}}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}})$ is guaranteed to be less than ε. In this case `EntGame` become automatically (ε, δ)-PAC.

It is equivalent to find a maximal T such that

$$\varepsilon T \leq 18 \log(SAT)\sqrt{H^4 SAT(\log(4SAH/\delta) + S\log(eT))\log(T)} + HSA\log(e(T+1))$$

and add 1 to it.

We start from obtaining a loose bound to eliminate logarithmic factors in T.

First, we assume that $T \geq 1$, thus $T + 1 \leq 2T$. Additionally, let us use inequality $\log(x) \leq x^\beta/\beta$ for any $x > 0$ and $\beta > 0$. We obtain

$$\varepsilon T \leq 18 \frac{(SAT)^{1/3}}{1/3}\sqrt{H^4 S^2 AT \log(4SAH/\delta)\frac{(eT)^{1/18}}{1/18}\frac{T^{1/18}}{1/18}} + HSA\frac{(2eT)^{8/9}}{8/9}$$

$$\leq T^{8/9}\Big(1010\sqrt{H^4 S^2 A^2 \log(2SAH/\delta)}\Big),$$

thus we can define $L = 9\log\Big(1010\sqrt{H^4 S^{8/3} A^{8/3} \log(4SAH/\delta)}/\varepsilon\Big)$ for which $\log(T) \leq L$. Thus we have

$$\varepsilon T \leq 18(\log(SA) + L)\sqrt{H^4 SAT(\log(4SAH/\delta) + S + L)L} + HSA(2 + L).$$

Solving this quadratic inequality, we obtain the minimal required T to guarantee $\mathcal{H}_{\text{visit}}(d^{\pi^{\star,\text{VE}}}) - \mathcal{H}_{\text{visit}}(d^{\widehat{\pi}}) \leq \varepsilon$. In particular,

$$T \geq 1 + \frac{648(\log(SA) + L)H^4 SA \cdot (\log(4SAH/\delta) + S + L) \cdot L}{\varepsilon^2} + \frac{2HSA(2 + L)}{\varepsilon}.$$

$\square$

C. Regularized Bellman Equations

In this section we provide complete proofs for regularized Bellman Equations in the general setting. Let $\Phi \colon \Delta_{\mathcal{A}} \to \mathbb{R}$ be a strictly convex function.

Then we can define the regularized value function as follows

$$V^{\pi}_{\lambda,h}(s) \triangleq \mathbb{E}_{\pi} \left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) - \lambda \Phi(\pi_{h'}(s_{h'})) \mid s_h = s \right]. \tag{7}$$

Notably, for a specific choice of rewards $r_h(s,a) = \mathcal{H}(p_h(s,a))$, the regularizer is equal to the negative entropy $\Phi(\pi) = -\mathcal{H}(\pi)$, and $\lambda = 1$ we have $V^{\pi}_{\lambda,1}(s_1) = \mathcal{H}_{\text{traj}}(q^{\pi})$, see Lemma H.2 In more general setting let $r_h(s,a)$ be equal to the sum of deterministic reward and $\lambda \mathcal{H}(p_h(s,a))$. In this case we have $V^{\pi}_{\lambda,1}(s_1) = V^{\pi}_1(s_1) + \lambda \mathcal{H}_{\text{traj}}(q^{\pi})$ in terms of a usual non-regularized value function.

Let us define an entropy action-value function as follows

$$Q^{\pi}_{\lambda,h}(s,a) \triangleq \mathbb{E}_{\pi} \left[r_h(s_h, a_h) + \sum_{h'=h+1}^{H} [r_{h'}(s_{h'}, a_{h'}) - \lambda \Phi(\pi_h(s_{h'}))] \mid (s_h, a_h) = (s,a) \right]. \tag{8}$$

Additionally, we define an optimal entropy-regularized value functions a follows

$$V^{\star}_{\lambda,h}(s) \triangleq \max_{\pi} V^{\pi}_{\lambda,h}(s), \quad Q^{\star}_{\lambda,h}(s,a) \triangleq \max_{\pi} Q^{\pi}_{\lambda,h}(s,a) \quad \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H].$$

C.1. Proof of Entropy-Regularized Bellman Equations

Theorem C.1 (Regularized Bellman Equations). *For any stochastic policy π the following decomposition of the entropy-regularized value function holds*

$$\begin{aligned} Q^{\pi}_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\pi}_{\lambda,h+1}(s,a), \\ V^{\pi}_{\lambda,h}(s) &= \pi_h Q^{\pi}_{\lambda,h}(s) - \lambda \Phi(\pi_h(s)). \end{aligned} \tag{9}$$

Moreover, for optimal Q- and V-functions we have

$$\begin{aligned} Q^{\star}_{\lambda,h}(s,a) &= r_h(s,a) + p_h V^{\star}_{\lambda,h+1}(s,a), \\ V^{\star}_{\lambda,h}(s) &= \max_{\pi \in \Delta_{\mathcal{A}}} \{ \pi Q^{\star}_h(s) - \lambda \Phi(\pi) \}. \end{aligned} \tag{10}$$

Remark C.2. For the case of interest $\Phi(\pi) = -\mathcal{H}(\pi)$ the expression for the V-function allows the closed-form formula by a well-known LogSumExp smooth maximum approximation

$$V^{\star}_{\lambda,h}(s) = \lambda \log \left(\sum_{a \in \mathcal{A}} \exp \left(\frac{1}{\lambda} Q^{\star}_{\lambda,h}(s,a) \right) \right),$$

and as $\lambda \to 0$ we see that entropy-regularized value function tends to a usual value function without regularization.

Proof. We proceed by induction. For $h = H + 1$ the equation is trivial. By definition and tower property of conditional expectation

$$\begin{aligned} Q^{\pi}_{\lambda,h}(s,a) &= r_h(s,a) + \mathbb{E} \left[\sum_{t=h+1}^{H} r_t(s_t, a_t) - \lambda \Phi(\pi_t(s_t)) \Big| s_h = s, a_h = a \right] \\ &= r(s,a) + \mathbb{E} \left[\mathbb{E} \left[\sum_{t=h+1}^{H} r_t(s_t, a_t) - \lambda \Phi(\pi_t(s_t)) \Big| s_{h+1} \right] \Big| s_h = s, a_h = a \right] \\ &= r_h(s,a) + p_h V^{\pi}_{\lambda,h+1}(s,a). \end{aligned}$$

Next we provide the second Bellman equation by tower property and the definition of the regularized Q-function

$$V^\pi_{\lambda,h}(s) = -\lambda\Phi(\pi_h(s)) + \mathbb{E}\left[r_h(s_h, a_h) + \sum_{t=h+1}^{H} r_t(s_t, a_t) - \lambda\Phi(\pi_t(s_t))\middle| s_h = s\right]$$

$$= \pi_h Q^\pi_{\lambda,h}(s) - \lambda\Phi(\pi_h(s)).$$

For optimal Bellman equation we proceed by induction. For $h = H + 1$ the equation is also trivial. By Bellman equations

$$Q^\star_{\lambda,h}(s, a) = \max_\pi\{r_h(s, a) + p_h V^\pi_{\lambda,h+1}(s, a)\} = r_h(s, a) + p_h V^\star_{\lambda,h+1}(s, a),$$

and, finally

$$V^\star_{\lambda,h}(s) = \max_{\pi_1,\ldots,\pi_H\in\Delta_\mathcal{A}}\{\pi_h Q^\star_{\lambda,h}(s) - \lambda\Phi(\pi_h(s))\} = \max_{\pi\in\Delta_\mathcal{A}}\{\pi Q^\star_{\lambda,h}(s) - \lambda\Phi(\pi)\}.$$

$\square$

C.2. A Bellman-type Equations for Variance

For a stochastic policy π we define Bellman-type equations for the variances as follows

$$\sigma Q^\pi_{\lambda,h}(s, a) \triangleq \mathrm{Var}_{p_h} V^\pi_{\lambda,h+1}(s, a) + p_h \sigma V^\pi_{\lambda,h+1}(s, a)$$

$$\sigma V^\pi_{\lambda,h}(s) \triangleq \mathrm{Var}_{\pi_h} Q^\pi_{\lambda,h}(s) + \pi_h \sigma Q^\pi_{\lambda,h}(s)$$

$$\sigma V^\pi_{\lambda,H+1}(s) \triangleq 0,$$

where $\mathrm{Var}_{p_h}(f)(s, a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[(f(s') - p_h f(s, a))^2]$ denotes the *variance operator over transitions* and $\mathrm{Var}_{\pi_h}(f)(s) \triangleq \mathbb{E}_{a'\sim\pi_h(s)}[(f(s, a') - \pi_h f(s))^2]$ denoted the *variance operator over the policy*. In particular, the function $s \mapsto \sigma V^\pi_{\lambda,1}(s)$ represents the average sum of the local variances $\mathrm{Var}_{p_h} V^\pi_{\lambda,h+1}(s, a)$ and $\mathrm{Var}_{\pi_h} Q^\pi_{\lambda,h}(s)$ over a trajectory following the policy π, starting from (s, a). Indeed, the definition above implies that

$$\sigma V^\pi_{\lambda,1}(s_1) = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}} d^\pi_h(s)\mathrm{Var}_{\pi_h} Q^\pi_{\lambda,h}(s) + \sum_{h=1}^{H}\sum_{s,a} d^\pi_h(s, a)\mathrm{Var}_{p_h}(V^\pi_{\lambda,h+1})(s, a).$$

The lemma below shows that we can relate the global variance of the cumulative reward over a trajectory to the average sum of local variances.

Lemma C.3 (Law of total variance). *For any stochastic policy π and for all $h \in [H]$,*

$$\sigma Q^\pi_{\lambda,h}(s, a) = \mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H}(r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - (Q^\pi_{\lambda,h}(s_h, a_h) - \lambda\Phi(\pi_h(s_h)))\right)^2\middle|(s_h, a_h) = (s, a)\right],$$

$$\sigma V^\pi_{\lambda,h}(s) = \mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H}(r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V^\pi_{\lambda,h}(s_h)\right)^2\middle|s_h = s\right].$$

Proof. We proceed by induction. The statement in Lemma C.3 is trivial for $h = H + 1$. We now assume that it holds for

$h + 1$ and prove that it also holds for h. For this purpose, we compute

$$\mathbb{E}_\pi \left[\left(\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) \right) - \left(Q_{\lambda,h}^\pi(s_h, a_h) - \lambda\Phi(\pi_h(s_h)) \right) \right)^2 \Bigg| (s_h, a_h) \right]$$

$$= \mathbb{E}_\pi \left[\left(V_{\lambda,h+1}^\pi(s_{h+1}) - p_h V_{\lambda,h+1}^\pi(s_h, a_h) + \sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h+1}^\pi(s_{h+1}) \right)^2 \Bigg| (s_h, a_h) \right]$$

$$= \mathbb{E}_\pi \left[\left(V_{\lambda,h+1}^\pi(s_{h+1}) - p_h V_{\lambda,h+1}^\pi(s_h, a_h) \right)^2 \Big| (s_h, a_h) \right]$$

$$+ \mathbb{E}_\pi \left[\left(\sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h+1}^\pi(s_{h+1}) \right)^2 \Bigg| (s_h, a_h) \right]$$

$$+ 2\mathbb{E}_\pi \left[\left(\sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h+1}^\pi(s_{h+1}) \right) \left(V_{\lambda,h+1}^\pi(s_{h+1}) - p_h V_{\lambda,h+1}^\pi(s_h, a_h) \right) \Bigg| (s_h, a_h) \right].$$

The definition of $V_{\lambda,h+1}^\pi(s_{h+1})$ implies that

$$\mathbb{E}_\pi \left[\sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h+1}^\pi(s_{h+1}) \Bigg| s_{h+1} \right] = 0.$$

Therefore, the tower property of conditional expectation gives us

$$\mathbb{E}_\pi \left[\left(\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) \right) - \left(Q_{\lambda,h}^\pi(s_h, a_h) - \lambda\Phi(\pi_h(s_h)) \right) \right)^2 \Bigg| (s_h, a_h) \right]$$

$$= \mathbb{E}_\pi \left[\left(V_{\lambda,h+1}^\pi(s_{h+1}) - p_h V_{\lambda,h+1}^\pi(s_h, a_h) \right)^2 \Big| (s_h, a_h) \right]$$

$$+ \mathbb{E} \left[\mathbb{E}_\pi \left[\left(\sum_{h'=h+1}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h+1}^\pi(s_{h+1}) \right)^2 \Bigg| s_{h+1} \right] \Bigg| (s_h, a_h) \right]$$

$$= \mathrm{Var}_{p_h} V_{\lambda,h+1}^\pi(s_h, a_h) + p_h \sigma V_{\lambda,h+1}^\pi(s_h, a_h) = \sigma Q_{\lambda,h}^\pi(s_h, a_h)$$

where in the third equality we used the inductive hypothesis and the definition of σV_{h+1}^π. To prove the second equation we use the entropy-regularized Bellman equations

$$\mathbb{E}_\pi \left[\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - V_{\lambda,h}^\pi(s_h) \right)^2 \Bigg| s_h = s \right]$$

$$= \mathbb{E}_\pi \left[\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - (Q_{\lambda,h}^\pi(s_h, a_h) - \lambda\Phi(\pi_h(s_h))) \right)^2 \Bigg| s_h = s \right]$$

$$+ 2\mathbb{E}_\pi \left[\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - (Q_{\lambda,h}^\pi(s_h, a_h) - \lambda\Phi(\pi_h(s_h))) \right) \left(\pi_h Q_{\lambda,h}^\pi(s_h) - Q_{\lambda,h}^\pi(s_h, a_h) \right) \Bigg| s_h = s \right]$$

$$+ \mathbb{E}_\pi \left[\left(\pi_h Q_{\lambda,h}^\pi(s_h) - Q_{\lambda,h}^\pi(s_h, a_h) \right)^2 \Big| s_h = s \right].$$

By definition of $Q_{\lambda,h}^\pi$ we have

$$\mathbb{E}_\pi \left[\left(\sum_{h'=h}^{H} (r_{h'}(s_{h'}, a_{h'}) - \lambda\Phi(\pi_{h'}(s_{h'}))) - (Q_{\lambda,h}^\pi(s_h, a_h) - \lambda\Phi(\pi_h(s_h))) \right) \Bigg| (s_h, a_h) = (s, a) \right] = 0,$$

thus, by the tower property

$$\mathbb{E}_{\pi}\left[\left(\sum_{h'=h}^{H}(r_{h'}(s_{h'},a_{h'})-\lambda\Phi(\pi_{h'}(s_{h'})))-V_{\lambda,h}^{\pi}(s_h)\right)^2 \Bigg| s_h = s\right] = \pi_h \sigma Q_{\lambda,h}^{\pi}(s_h) + \mathrm{Var}_{\pi_h}Q_{\lambda,h}^{\pi}(s_h) = \sigma V_{\lambda,h}^{\pi}(s_h).$$

$\square$

C.3. Performance-Difference Lemma

In this section we provide a version of performance-difference lemma (see e.g. Lemma E.15 by Dann et al. (2017)) for regularized Bellman equations.

Lemma C.4 (Performance-Difference Lemma). *Let* $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, H, \{p'_h\}_{h\in[H]}, \{r'_h\}_{h\in[H]}, s_1)$ *and* $\mathcal{M}'' = (\mathcal{S}, \mathcal{A}, H, \{p''_h\}_{h\in[H]}, \{r''_h\}_{h\in[H]}, s_1)$ *be two MDPs and let* $Q_{\lambda,h}^{\mathcal{M},\pi}(s,a)$ *be a Q-value of policy* π *in the MDP* $\mathcal{M}$ *with regularization. Then for any* $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$,

$$Q_{\lambda,h}^{\mathcal{M}',\pi}(s,a) - Q_{\lambda,h}^{\mathcal{M}'',\pi}(s,a) = \mathbb{E}_{\mathcal{M}'',\pi}\left[\sum_{h'=h}^{H} r'_{h'}(s_{h'},a_{h'}) - r''_{h'}(s_{h'},a_{h'}) \,\Bigg|\, (s_h,a_h) = (s,a)\right]$$

$$+ \mathbb{E}_{\mathcal{M}'',\pi}\left[\sum_{h'=h}^{H}[p'_{h'} - p''_{h'}]V_{\lambda,h'+1}^{\mathcal{M}',\pi}(s_{h'},a_{h'}) \,\Bigg|\, (s_h,a_h) = (s,a)\right].$$

Proof. Let us proceed by induction over h. For $h = H + 1$ this statement is trivially true. Next we assume that it holds for any $h' > h$. By regularized Bellman equations

$$Q_{\lambda,h}^{\mathcal{M}',\pi}(s,a) - Q_{\lambda,h}^{\mathcal{M}'',\pi}(s,a) = [r'_h(s,a) - r''_h(s,a)] + p'_h V_{\lambda,h}^{\mathcal{M}',\pi}(s,a) - p''_h V_{\lambda,h}^{\mathcal{M}'',\pi}(s,a)$$

$$= [r'_h(s,a) - r''_h(s,a)] + [p'_h - p''_h]V_{\lambda,h+1}^{\mathcal{M}',\pi}(s,a) - p''_h\left[V_{\lambda,h+1}^{\mathcal{M}'',\pi} - V_{\lambda,h+1}^{\mathcal{M}',\pi}\right](s,a)$$

$$= [r'_h(s,a) - r''_h(s,a)] + [p'_h - p''_h]V_{\lambda,h+1}^{\mathcal{M}',\pi}(s,a) + p''_h\left[V_{\lambda,h+1}^{\mathcal{M}',\pi} - V_{\lambda,h+1}^{\mathcal{M}'',\pi}\right](s,a).$$

Next we notice that

$$V_{\lambda,h+1}^{\mathcal{M}',\pi}(s) - V_{\lambda,h+1}^{\mathcal{M}'',\pi}(s) = \pi Q_{\lambda,h+1}^{\mathcal{M}',\pi}(s) - \lambda\Phi(\pi) - \pi Q_{\lambda,h+1}^{\mathcal{M}'',\pi}(s) + \lambda\Phi(\pi) = \pi\left[Q_{\lambda,h+1}^{\mathcal{M}',\pi} - Q_{\lambda,h+1}^{\mathcal{M}'',\pi}\right](s)$$

since the regularizer is cancelled out. Thus, we can rewrite this difference as follows

$$Q_{\lambda,h}^{\mathcal{M}',\pi}(s,a) - Q_{\lambda,h}^{\mathcal{M}'',\pi}(s,a) = \mathbb{E}_{\pi,\mathcal{M}''}\left[r'_h(s_h,a_h) - r''_h(s_h,a_h) + [p'_h - p''_h]V_{\lambda,h+1}^{\mathcal{M}',\pi}(s_h,a_h)\right.$$

$$\left. + \left[Q_{\lambda,h+1}^{\mathcal{M}',\pi}(s_{h+1},a_{h+1}) - Q_{\lambda,h+1}^{\mathcal{M}'',\pi}(s_{h+1},a_{h+1})\right]\,\Bigg|\,(s_h,a_h) = (s,a)\right].$$

By induction hypothesis we conclude the statement. $\square$

D. Sample Complexity for MTEE and Regularized MDPs

In this section we describe the general setting of regularized MDPs, not only entropy-regularized.

D.1. Preliminaries

First we define class of regularizers we are interested in. For more exposition on this definition, see Bubeck (2015).

Definition D.1. Let $\Phi\colon \Delta_{\mathcal{A}} \to \mathbb{R}$ be a proper closed strongly-convex function. We will call Φ a mirror-map if the following holds

- Φ is 1-strongly convex with respect to norm $\|\cdot\|$;

- $\nabla\Phi$ takes all possible values in $\mathbb{R}^{\mathcal{A}}$;

- $\nabla\Phi$ diverges on the boundary of $\Delta_{\mathcal{A}}$: $\lim_{x\in\partial\Delta_{\mathcal{A}}}\|\nabla\Phi(x)\| = +\infty$;

We explain three main examples of a such regularizers.

- The negative Shannon entropy $\Phi(\pi) = -\mathcal{H}(\pi)$ for $\mathcal{H}(\pi) = \sum_{a\in\mathcal{A}} \pi_a \log\left(\frac{1}{\pi_a}\right)$ satisfies the Definition D.1 for ℓ_1-norm;

- The negative Tsallis entropy $\Phi(\pi) = -\frac{1}{q}\mathcal{T}_q(\pi)$ for $\mathcal{T}_q(\pi) = \frac{1}{q-1}\left(1 - \sum_{a\in\mathcal{A}} \pi_a^q\right)$ satisfied the Definition D.1 for ℓ_2 norm for every $q \in (0,1)$. In particular, $q = 0.5$ corresponds to the choice by Grill et al. (2019) in Appendix E that is tightly connected to the UCB algorithm;

- For any other fixed policy $\pi' \in \Delta_{\mathcal{A}}$ we can choose $\Phi(\pi) = \mathrm{KL}(\pi, \pi') = \sum_{a\in\mathcal{A}} \pi_a \log\left(\frac{\pi_a}{\pi'_a}\right)$ that inherits all the properties from the choice of the negative entropy.

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}, s_1)$ be a finite-horizon MDP, where $r_h(s,a)$ is a deterministic reward function. For simplicity we assume that $0 \le r_h(s,a) \le r_{\max}$ for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$.

Then we can define entropy-augmented rewards as follows

$$r_{\kappa,h}(s,a) = r_h(s,a) + \kappa\mathcal{H}(p_h(s,a)).$$

This definition is required to cover the following case of practical interest. Let $\kappa = \lambda$ and $\Phi(\pi) = -\mathcal{H}(\pi)$, then we obtain the following representation for the λ-regularized value function

$$V_{\lambda,1}^{\pi}(s_1) = V_1^{\pi}(s_1) + \lambda\mathcal{H}_{\mathrm{traj}}(q^{\pi}),$$

where $V_1^{\pi}(s_1)$ is a usual value function for a MDP $\mathcal{M}$. For $r_{\max} = 0$ and $\kappa = \lambda = 1$ we recover just a trajectory entropy $V_{\lambda,1}^{\pi}(s_1) = \mathcal{H}_{\mathrm{traj}}(q^{\pi})$.

Next we define a convex conjugate to $\lambda\Phi$ as $F_\lambda\colon \mathbb{R}^{\mathcal{A}} \to \mathbb{R}$

$$F_\lambda(x) = \max_{\pi\in\Delta_{\mathcal{A}}}\{\langle\pi, x\rangle - \lambda\Phi(\pi)\}$$

and, with a sight abuse of notation extend the action of this function to the Q-function as follows

$$V_{\lambda,h}^{\star}(s) = F_\lambda(Q_{\lambda,h}^{\star})(s) = \max_{\pi\in\Delta_{\mathcal{A}}}\{\pi Q_{\lambda,h}^{\star}(s) - \lambda\Phi(\pi)\}.$$

Thanks to the fact that Φ satisfies Definition D.1, we have exact formula for the optimal policy by Fenchel-Legendre transform

$$\pi_h^{\star}(s) = \arg\max_{\pi\in\Delta_{\mathcal{A}}}\{\pi Q_{\lambda,h}^{\star}(s) - \lambda\Phi(\pi)\} = \nabla F_\lambda(Q_{\lambda,h}^{\star}(s,\cdot)).$$

Notice that we have $\nabla F_\lambda(Q_{\lambda,h}^{\star}(s,\cdot)) \in \Delta_{\mathcal{A}}$ since the gradient of Φ diverges on the boundary of $\Delta_{\mathcal{A}}$. For entropy regularization this formula become the softmax function.

Finally, it is known that the smoothness property of F_λ plays a key role in reduced sample complexity for planning in regularized MDPs (Grill et al., 2019). For our general setting we have that since $\lambda\Phi$ is λ-strongly convex with respect to $\|\cdot\|$, then F_λ is $1/\lambda$-strongly smooth with respect to a dual norm $\|\cdot\|_*$

$$F_\lambda(x) \leq F_\lambda(x') + \langle \nabla F_\lambda(x'), x - x' \rangle + \frac{1}{2\lambda}\|x - x'\|_*^2.$$

Let us define R_Φ as a maximal possible value of $|\Phi|$. Without loss of generality assume that $\Phi \leq 0$. In this case we define $R_{\max} = r_{\max} + \kappa \log(S) + \lambda R_\Phi$ as an upper bound of an about of reward obtain at the one step. By this definition we have $0 \leq V_{\lambda,h}^\pi(s) \leq HR_{\max}$ for any $h \in [H], s \in \mathcal{S}$ and any policy π.

Also, since all norms in $\mathbb{R}^{\mathcal{A}}$ are equivalent, we define a constant r_A that defined for a dual norm $\|\cdot\|_*$ as follows

$$\|\cdot\|_* \leq r_A \cdot \|\cdot\|_\infty. \tag{11}$$

For example, for ℓ_2-norm $r_A = \sqrt{A}$ and for ℓ_1-norm $r_A = A$. In the case $\Phi = -\mathcal{H}$ we have $r_A = 1$ since the entropy is 1-strongly convex with respect to a ℓ_1-norm, thus the dual norm is exactly a ℓ_∞-norm.

The rest of this section is devoted to obtain the sample complexity guarantees for `UCBVI-Ent` algorithm with *a regularization-agnostic stopping rule* (17) with the gap notion (16). In this case Theorem D.9 gives us $\widetilde{\mathcal{O}}\left(\frac{H^3 SA}{\varepsilon^2} + \frac{H^3 S^2 A}{\varepsilon} \right)$ sample complexity guarantee ignoring $R_{\max}$ and poly-logarithmic factors. Notably, this sample complexity result does depend directly on $1/\lambda$, so small regularization does not affect the complexity.

In Section E we present another algorithm `RL-Explore-Ent`, based on ideas of reward-free exploration, that achieves sample complexity of order $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/(\lambda\varepsilon)$. As a particular example, it yields an algorithm for the MTEE problem with sample complexity $\widetilde{\mathcal{O}}(\mathrm{poly}(S, A, H)/\varepsilon)$ by taking as a regularizer negative entropy, $\lambda = \kappa = 1$ and $r_{\max} = 0$. Moreover, in Section F we apply this algorithm to achieve sample complexity of order $\widetilde{\mathcal{O}}(H^2 SA/\varepsilon^2)$ for the maximum visitation entropy exploration problem.

D.2. `UCBVI-Ent` Algorithm

In this section we describe `UCBVI-Ent` algorithm to solve MTEE problem, however as we show in the proofs, it is capable to work with general regularized MDPs.

We now describe our algorithm `UCBVI-Ent` for MTEE. Since one only needs to solve regularized Bellman equations to obtain a maximum trajectory entropy policy, we can use an algorithm of the same flavor as the ones proposed for best policy identification (Tirinzoni et al., 2023; Ménard et al., 2021; Kaufmann et al., 2021; Dann et al., 2019). In particular, `UCBVI-Ent` is close to the `BPI-UCRL` algorithm by Kaufmann et al. (2021) and can be characterized by the following rules.

Sampling rule   As sampling rule we use an optimistic policy π^{t+1} obtained by optimistic planning in the regularized MDP

$$\overline{Q}_h^t(s, a) = \mathrm{clip}\left(\mathcal{H}\big(\widehat{p}_h(s, a)\big) + b_h^{\mathcal{H},t}(s, a) + \widehat{p}_h^t \overline{V}_{h+1}^t(s, a) + b_h^{p,t}(s, a), 0, \log(SA)H \right),$$

$$\overline{V}_h^t(s) = \max_{\pi \in \Delta_A} \pi \overline{Q}_h^t(s) + \mathcal{H}(\pi), \tag{12}$$

$$\pi_h^{t+1}(s) = \arg\max_{\pi \in \Delta_A} \pi \overline{Q}_h^t(s) + \mathcal{H}(\pi),$$

with $\overline{V}_{H+1}^t = 0$ by convention where $b^{\mathcal{H},t}, b^{p,t}$ are bonuses for the entropy and the transition probabilities, respectively. Precisely, we use bonuses of the form

$$b_h^{\mathcal{H},t}(s, a) = \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s, a))}{n_h^t(s, a)}} + \min\left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s, a))}{n_h^t(s, a)}, \log(S) \right),$$

$$b_h^{p,t}(s, a) \triangleq b_h^{B,t}(s, a) + b_h^{\mathrm{corr},t}(s, a),$$

$$b_h^{B,t}(s, a) \triangleq 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\overline{V}_{h+1}^t)(s, a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s, a))}{n_h^t(s, a)}} + \frac{9H^2 \log(SA)\beta^{\mathrm{KL}}(\delta, n_h^t(s, a))}{n_h^t(s, a)},$$

$$b_h^{\mathrm{corr},t}(s, a) \triangleq \frac{1}{H}\widehat{p}_h^t(\overline{V}_{h+1}^t - \underline{V}_{h+1}^t)(s, a).$$

for some functions $\beta^{\mathcal{H}}, \beta^{\mathrm{KL}}$ and β^{conc} and $\underline{V}^t$ is a lower confidence bound on the optimal value function defined in Appendix D.4.

Stopping and decision rule To define the stopping rule, we first recursively build an upper-bound on the difference between the value of the optimal policy and the value of the current policy π^{t+1},

$$
G_h^t(s,a) \triangleq \mathrm{clip}\bigg(2b_h^{B,t}(s,a) + 2b_h^{\mathcal{H},t}(s,a) + \frac{4H^2 \log(SA) \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \\
+ \Big(1 + \frac{3}{H}\Big) \widehat{p}_h^t \big[\pi_{h+1}^{t+1} G_{h+1}^t\big](s,a), 0, H\mathrm{R}_{\max} \bigg),
\tag{13}
$$

where $\underline{V}^t$ is a lower-bound on the optimal value function defined in Appendix D.4 and $G_{H+1}^t(s,a) = 0$ by convention. Then the stopping time $\tau = \inf\{t \in \mathbb{N} : \pi^{t+1} G_1^t(s_1) \leq \varepsilon\}$ corresponds to the first episode when this upper-bound is smaller than ε. At this episode we return the Markovian policy $\widehat{\pi} = \pi^{\tau+1}$.

Algorithm 3 `UCBVI-Ent`

1: **Input:** Target precision ε, target probability δ, threshold functions $\beta^{\mathcal{H}}, \beta^p, \beta^{\mathrm{conc}}$.
2: **while** true **do**
3:    Compute π^t by optimistic planning with (12).
4:    Compute bound on the gap $G_1^{t-1}(s,a)$ with (13).
5:    **if** $\pi^t G_1^{t-1}(s_1) \leq \varepsilon$ **then break**
6:    **for** $h \in [H]$ **do**
7:        Play $a_h^t \sim \pi_h^t(s_h^t)$
8:        Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
9:    **end for**
10:   Update counts n^t, transition estimates $\widehat{p}^t$ and episode number $t \leftarrow t+1$.
11: **end while**
12: Output policy $\widehat{\pi} = \pi^t$.

The complete procedure is described in Algorithm 3. We now state our main theoretical result for `UCBVI-Ent`. We prove that for the well calibrated threshold functions $\beta^{\mathcal{H}}, \beta^{\mathrm{KL}}$ and β^{conc}, the `UCBVI-Ent` is (ε, δ)-PAC for MTEE and provide a high-probability upper bound on its sample complexity.

Theorem D.2. *Let $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}$ and $\beta^{\mathcal{H}}$ be defined in Lemma D.3 of Appendix D. Fix $\varepsilon > 0$ and $\delta \in (0,1)$, then the* `UCBVI-Ent` *algorithm is (ε, δ)-PAC for MTEE. Moreover, the optimal policy is given by $\widehat{\pi} = \pi^{\tau+1}$ where*

$$
\tau = \widetilde{\mathcal{O}}\bigg(\frac{H^3 SA}{\varepsilon^2} + \frac{H^3 S^2 A}{\varepsilon} \bigg).
$$

with probability at least $1 - \delta$. Here $\widetilde{\mathcal{O}}$ hides poly-logarithmic factors in $\varepsilon, \delta, H, S, A$.

See Theorem D.10 in Appendix D for a precise bound and a proof. Basically, this result is a simple corollary of the general result on regularized MDPs.

Space and time complexity Since `UCBVI-Ent` is a model based algorithm, its space-complexity is of order $\mathcal{O}(HS^2A)$ whereas its time-complexity for one episode is of order $\mathcal{O}(HS^2A)$ because of the optimistic planning.

D.3. Concentration Events

Following the ideas of (Ménard et al., 2021), we define the following concentration events.

Let $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathrm{cnt}}, \beta^{\mathcal{H}} : (0,1) \times \mathbb{N} \to \mathbb{R}_+$ be some functions defined later on in Lemma D.3. We define the following

favorable events

$$\mathcal{E}^{\mathrm{KL}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leq \frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \right\},$$

$$\mathcal{E}^{\mathrm{conc}}(\delta) \triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}:$$

$$|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \leq \sqrt{2\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 3H\mathrm{R}_{\max}\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \Bigg\},$$

$$\mathcal{E}^{\mathrm{cnt}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad n_h^t(s,a) \geq \frac{1}{2}\overline{n}_h^t(s,a) - \beta^{\mathrm{cnt}}(\delta) \right\},$$

$$\mathcal{E}^{\mathcal{H}}(\delta) \triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}:$$

$$|\mathcal{H}(\widehat{p}_h^t(s,a)) - \mathcal{H}(p_h(s,a))| \leq \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \left(\frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge \log(S) \right) \Bigg\}.$$

We also introduce two intersections of these events of interest, $\mathcal{G}(\delta) \triangleq \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}_B^{\mathrm{conc}}(\delta) \cap \mathcal{E}^{\mathrm{cnt}}(\delta) \cap \mathcal{E}^{\mathcal{H}}(\delta)$. We prove that for the right choice of the functions $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathrm{cnt}}, \beta^{\mathcal{H}}$ the above events hold with high probability.

Lemma D.3. *For any $\delta \in (0,1)$ and for the following choices of functions β,*

$$\beta^{\mathrm{KL}}(\delta, n) \triangleq \log(4SAH/\delta) + S\log(\mathrm{e}(1+n)),$$
$$\beta^{\mathrm{conc}}(\delta, n) \triangleq \log(4SAH/\delta) + \log(4\mathrm{e}n(2n+1)),$$
$$\beta^{\mathrm{cnt}}(\delta) \triangleq \log(4SAH/\delta),$$
$$\beta^{\mathcal{H}}(\delta, n) \triangleq \log^2(n)(\log(4SAH/\delta) + \log(n(n+1))),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1 - \delta/4, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geq 1 - \delta/4,$$
$$\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geq 1 - \delta/4, \qquad \mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geq 1 - \delta/4.$$

In particular, $\mathbb{P}[\mathcal{G}(\delta)] \geq 1 - \delta$.

Proof. Applying Theorem G.1 and the union bound over $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1 - \delta/4$. Next, Theorem G.4 and the union bound over $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$ yield $\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geq 1 - \delta/4$. By Theorem G.3 and union bound, $\mathbb{P}[\mathcal{E}^{\mathrm{cnt}}(\delta)] \geq 1 - \delta/4$. Finally, by Theorem G.2 and union bound over $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$ $\mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geq 1 - \delta/4$. The union bound over four prescribed events concludes $\mathbb{P}[\mathcal{G}(\delta)] \geq 1 - \delta$. $\square$

Lemma D.4. *Assume conditions of Lemma D.3. Then on event $\mathcal{E}^{\mathrm{KL}}(\delta)$, for any $f: \mathcal{S} \to [0, H\mathrm{R}_{\max}], t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$[p_h - \widehat{p}_h^t]f(s,a) \leq \frac{1}{H}\widehat{p}_h^t f(s,a) + 2H\mathrm{R}_{\max}\left(\frac{2H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1 \right),$$

$$[\widehat{p}_h^t - p_h]f(s,a) \leq \frac{1}{H}p_h f(s,a) + 2H\mathrm{R}_{\max}\left(\frac{2H\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1 \right).$$

Proof. Let us start from the first statement. We apply the second inequality of Lemma H.7 and Lemma H.8 to obtain

$$[p_h - \widehat{p}_h^t]f(s,a) \leq \sqrt{2\mathrm{Var}_{p_h}[f](s,a) \cdot \mathrm{KL}(\widehat{p}_h^t, p_h)}$$
$$\leq 2\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[f](s,a) \cdot \mathrm{KL}(\widehat{p}_h^t, p_h)} + 3H\mathrm{R}_{\max}\mathrm{KL}(\widehat{p}_h^t, p_h).$$

Since $0 \le f(s) \le HR_{\max}$ we get

$$\mathrm{Var}_{\widehat{p}_h^t}[f](s,a) \le \widehat{p}_h^t[f^2](s,a) \le HR_{\max} \cdot \widehat{p}_h^t f(s,a).$$

Finally, applying $2\sqrt{ab} \le a + b, a, b \ge 0$, we obtain the following inequality

$$(\widehat{p}_h^t - p_h)f(s,a) \le \frac{1}{H}\widehat{p}_h^t f(s,a) + 4H^2 R_{\max} \mathrm{KL}(\widehat{p}_h^t, p_h).$$

Definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$ implies the part of the statement. At the same time we have a trivial bound since $f(s) \in [0, HR_{\max}]$

$$[p_h - \widehat{p}_h^t]f(s,a) \le 2HR_{\max} \le \frac{1}{H}\widehat{p}_h^t f(s,a) + 2HR_{\max}.$$

To prove the second statement, apply the first inequality of Lemma H.7 and proceed similarly. $\qquad\square$

Lemma D.5. *Assume conditions of Lemma D.3 and assume that $\beta^{\mathrm{conc}}(\delta) \le \beta^{\mathrm{KL}}(\delta)$. Then conditioned on event $\mathcal{G}(\delta)$, for any $U \colon \mathcal{S} \to [0, HR_{\max}]$, $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \le 3\sqrt{\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \frac{9H^2 R_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}$$

$$+ \frac{1}{H}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a).$$

Proof. First, we apply the definition of event $\mathcal{E}^{\mathrm{conc}}(\delta)$

$$|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \le \sqrt{2\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 3HR_{\max}\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Next we apply Lemma H.8 and Lemma H.9 and obtain

$$\mathrm{Var}_{p_h}(V_{\lambda,h+1}^\star)(s,a) \le 2\mathrm{Var}_{\widehat{p}_h^t}(V_{\lambda,h+1}^\star)(s,a) + 4H^2 R_{\max}^2 \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a))$$

$$\le 4\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a) + 4HR_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) + 4H^2 R_{\max}^2 \mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)).$$

Thus, by inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$.

$$|(\widehat{p}_h^t - p_h)V_{\lambda,h+1}^\star(s,a)| \le 3\sqrt{\mathrm{Var}_{\widehat{p}_{h+1}^t}(U)(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ 3\sqrt{HR_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) \cdot \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ 3HR_{\max}\sqrt{\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \cdot \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ 3HR_{\max}\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

By inequality $2\sqrt{ab} \le a + b$ we have

$$3\sqrt{HR_{\max}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) \cdot \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \le \frac{1}{H}\widehat{p}_h^t|U - V_{\lambda,h+1}^\star|(s,a) + \frac{9H^2 R_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{4n_h^t(s,a)}.$$

By the definition of the event $\mathcal{E}^{\mathrm{KL}}(\delta)$ and the fact $\beta^{\mathrm{conc}}(\delta) \le \beta^{\mathrm{KL}}(\delta)$ we have

$$\sqrt{\mathrm{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \cdot \frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \le \frac{\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

$\qquad\square$

D.4. Confidence Intervals

Similar to Azar et al. (2017); Zanette & Brunskill (2019); Ménard et al. (2021), we define the upper confidence bound for the optimal regularized Q-function of two types: with Hoeffding bonuses and with Bernstein bonuses.

Let us define empirical estimate of entropy-augmented rewards as follows

$$\hat{r}^t_{\kappa,h}(s,a) = r_h(s,a) + \kappa \mathcal{H}(\hat{p}^t_h(s,a)).$$

Then we have the following sequences defined as follows

$$\overline{Q}^t_h(s,a) = \text{clip}\Big(\hat{r}^t_{\kappa,h}(s,a) + \hat{p}^t_h \overline{V}^t_h(s,a) + b^{p,t}_h(s,a) + \kappa b^{\mathcal{H},t}_h(s,a), 0, H\text{R}_{\max}\Big)$$

$$\pi^{t+1}_h(s) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\pi \overline{Q}^t_h(s) - \lambda \Phi(\pi)\},$$

$$\overline{V}^t_h(s) = \mathcal{H}(\pi^{t+1}_h(s)) + \pi^{t+1}_h \overline{Q}^t_h(s)$$

$$\overline{V}^t_{H+1}(s) = 0,$$

and the lower confidence bound as follows

$$\underline{Q}^t_h(s,a) = \text{clip}\Big(\hat{r}^t_{\kappa,h}(s,a) + \hat{p}^t_h \underline{V}^t_h(s,a) - b^{p,t}_h(s,a) - \kappa b^{\mathcal{H},t}_h(s,a), 0, H\text{R}_{\max}\Big)$$

$$\underline{V}^t_h(s) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\pi \underline{Q}^t_h(s) - \lambda \Phi(\pi)\}$$

$$\underline{V}^t_{H+1}(s) = 0,$$

where the Bernstein bonuses for transitions are defined as follows

$$b^{p,t}_h(s,a) \triangleq b^{B,t}_h(s,a) + b^{\text{corr},t}_h(s,a),$$

$$b^{B,t}_h(s,a) \triangleq 3\sqrt{\text{Var}_{\hat{p}^t_h}(\overline{V}^t_{h+1})(s,a) \frac{\beta^{\text{conc}}(\delta, n^t_h(s,a))}{n^t_h(s,a)}} + \frac{9H^2\text{R}_{\max}\beta^{\text{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)}, \tag{14}$$

$$b^{\text{corr},t}_h(s,a) \triangleq \frac{1}{H}\hat{p}^t_h(\overline{V}^t_{h+1} - \underline{V}^t_{h+1})(s,a).$$

The entropy bonuses are defined bellow

$$b^{\mathcal{H},t}_h(s,a) \triangleq \sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n^t_h(s,a))}{n^t_h(s,a)}} + \left(\frac{\beta^{\text{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)} \wedge \log(S)\right). \tag{15}$$

Theorem D.6. *Let $\delta \in (0,1)$. Assume Bernstein bonuses* (14). *Then on event $\mathcal{G}(\delta)$ for any $t \in \mathbb{N}$, $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$ it holds*

$$\underline{Q}^t_h(s,a) \leq Q^\star_{\lambda,h}(s,a) \leq \overline{Q}^t_h(s,a), \qquad \underline{V}^t_h(s) \leq V^\star_{\lambda,h}(s) \leq \overline{V}^t_h(s,a).$$

Proof. Proceed by induction over h. For $h = H+1$ the statement is trivial. Now we assume that inequality holds for any $h' > h$ for a fixed $h \in [H]$. Fix a timestamp $t \in \mathbb{N}$ and a state-action pair (s,a) and assume that $\overline{Q}^t_h(s,a) < H\text{R}_{\max}$, i.e. no clipping occurs. Otherwise the inequality $Q^\star_{\lambda,h}(s,a) \leq \overline{Q}^t_h(s,a)$ is trivial. In particular, it implies $n^t_h(s,a) > 0$.

In this case, by Bellman equations (10) we have

$$[\overline{Q}^t_h - Q^\star_{\lambda,h}](s,a) = \underbrace{r_h(s,a) + \kappa\mathcal{H}(\hat{p}^t_h(s,a)) - r_h(s,a) - \kappa\mathcal{H}(p_h(s,a)) + \kappa b^{\mathcal{H},t}_h(s,a)}_{T_1}$$

$$+ \underbrace{\hat{p}^t_h \overline{V}^t_{h+1}(s,a) - p_h V^\star_{\lambda,h+1}(s,a) + b^{p,t}_h(s,a)}_{T_2}.$$

By the definition of event $\mathcal{E}^{\mathcal{H}}(\delta)$ that is subset of $\mathcal{G}(\delta)$ we have $T_1 \geq 0$. To show that $T_2 \geq 0$, we start from induction hypothesis

$$T_2 \geq [\hat{p}^t_h - p_h]V^\star_{\lambda,h+1}(s,a) + b^{p,t}_h(s,a).$$

Next we apply Lemma D.5 with $U = \overline{V}_{h+1}^t$ and definition of transition bonuses we have

$$T_2 \geq -\frac{1}{H}\widehat{p}_h^t|\overline{V}_{h+1}^t - V_{\lambda,h+1}^\star|(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

By induction hypothesis we have $\overline{V}_{h+1}^t(s) \geq V_{\lambda,h+1}^\star(s) \geq \underline{V}_{h+1}^t(s)$, thus $T_2 \geq 0$.

To prove the second inequality on Q-function, we assume $\underline{Q}_h^t(s,a) > 0$ and, as a consequence, $n_h^t(s,a) > 0$. Thus we have

$$[\underline{Q}_h^t - Q_{\lambda,h}^\star](s,a) = \underbrace{\mathcal{H}(\widehat{p}_h^t(s,a)) - \mathcal{H}(p_h(s,a)) - b_h^{\mathcal{H},t}(s,a)}_{T_1'} + \underbrace{\widehat{p}_h^t\underline{V}_{h+1}^t(s,a) - p_hV_{\lambda,h+1}^\star(s,a) - b_h^{p,t}(s,a)}_{T_2'}.$$

Again, by the definition of event $\mathcal{E}^{\mathcal{H}}(\delta)$ we have $T_1' \leq 0$ and, by induction hypothesis

$$T_2' \leq [\widehat{p}_h^t - p_h]V_{h+1}^{\mathcal{H},\star}(s,a) - b_h^{p,t}(s,a).$$

We again apply Lemma D.5 with $U = \overline{V}_{h+1}^t$

$$T_2' \leq \frac{1}{H}\widehat{p}_h^t|\overline{V}_{h+1}^t - V_{\lambda,h+1}^\star|(s,a) - \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

We conclude the statement by induction hypothesis for $h' = h + 1$.

Finally, we have to show the inequality for V-functions. To do it, we use the fact that V-functions are computed by F_λ applied to Q-functions

$$\underline{V}_h^t(s) = F_\lambda(\underline{Q}_h^t)(s), \quad V_{\lambda,h}^\star(s) = F_\lambda(Q_{\lambda,h}^\star)(s), \quad \overline{V}_h^t(s) = F_\lambda(\overline{Q}_h^t)(s).$$

Notice that ∇F_λ takes values in a probability simplex, thus, all partial derivatives of F_λ are non-negative and therefore F_λ is monotone in each coordinate. Thus, since $\underline{Q}_h^t(s,a) \leq Q_{\lambda,h}^\star(s,a) \leq \overline{Q}_h^t(s,a)$, we have the same inequality $\underline{V}_h^t(s) \leq V_{\lambda,h}^\star(s) \leq \overline{V}_h^t(s)$. □

D.5. Regularization-Agnostic Stopping Rule

In this section we provide guarantees for the so-called *regularization-agnostic gap*: this notion of gap does not influenced by regularization except the changing of the range of value functions and basically mimics UCBVI-BPI algorithm by Ménard et al. (2021) in definition of the similar gap.

Let us define $G_{H+1}^t(s,a) \triangleq 0$ for all s, a and

$$G_h^t(s,a) \triangleq \text{clip}\bigg(2b_h^{B,t}(s,a) + \frac{4H^2\text{R}_{\max}\beta^{\text{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + 2\kappa b_h^{\mathcal{H},t}(s,a) + \bigg(1 + \frac{3}{H}\bigg)\widehat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a),$$
$$0, H\text{R}_{\max}\bigg), \tag{16}$$

where $b_h^{B,t}(s,a)$ is defined in (14). For this notion of the gap we can define the stopping rule as follows

$$\tau = \min\{t \in \mathbb{N} : \pi_1^{t+1}G_1^t(s_1) \leq \varepsilon\}. \tag{17}$$

The lemma below justifies this choice of the stopping time.

Lemma D.7. *Assume the choice of Bernstein bonuses* (14) *and let the event $\mathcal{G}(\delta)$ defined in Lemma D.3 holds. Then for all $t \in \mathbb{N}$ we have*

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) \leq \pi_1^{t+1}G_1^t(s_1).$$

Proof. Following (Ménard et al., 2021), we start by defining the following quantities

$$\tilde{Q}_h^t(s,a) \triangleq \text{clip}\Big(\hat{r}_{\kappa,h}^t(s,a) + \hat{p}_h^t \tilde{V}_{h+1}^t(s,a) - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a), 0, r_{\kappa,h}(s,a) + p_h \tilde{V}_{h+1}^t(s,a)\Big),$$

$$\tilde{V}_h^t(s) \triangleq \pi_h^{t+1} \tilde{Q}_h^t(s) - \lambda \Phi(\pi_h^{t+1}(s)),$$

$$\tilde{V}_{H+1}^t(s) \triangleq 0.$$

By Theorem D.6 and Lemma D.8 we have

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) \leq \overline{V}_1^t(s_1) - V_{\lambda,1}^{\pi^{t+1}}(s_1) \leq \overline{V}_1^t(s_1) - \tilde{V}_1^t(s_1)$$

$$= \pi_1^{t+1}\overline{Q}_1^t(s_1) - \lambda\Phi(\pi_1^{t+1}(s_1)) - \pi_1^{t+1}\tilde{Q}_1^t(s_1) + \lambda\Phi(\pi_1^{t+1}(s_1)) = \pi_1^{t+1}[\overline{Q}_1^t - \tilde{Q}_1^t](s_1).$$

Therefore, it is enough to show that for any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$

$$[\overline{Q}_h^t - \tilde{Q}_h^t](s,a) \leq G_h^t(s,a), \qquad [\overline{V}_h^t - \tilde{V}_h^t](s) \leq \pi_h^{t+1} G_h^t(s).$$

Proceed by backward induction over h. The case $h = H + 1$ is trivial, thus we may assume that the statement holds for any $h' > h$ for a fixed h. Also fix $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Notice that if $G_h^t(s,a) = H\text{R}_{\max}$, then the inequality is trivially true. Therefore we may assume that $G_h^t(s,a) < H\text{R}_{\max}$ and, consequently, $n_h^t(s,a) > 0$. Now we have to separate cases.

First case. In this case we have $\tilde{Q}_h^t(s,a) = r_{\kappa,h}(s,a) + p_h \tilde{V}_{h+1}^t(s,a)$, i.e. maximal clipping occurs. Therefore

$$\overline{Q}_h^t(s,a) - \tilde{Q}_h^t(s,a) = r_h(s,a) + \kappa\mathcal{H}(\hat{p}_h^t(s,a)) + \kappa b_h^{\mathcal{H},t}(s,a) - r_h(s,a) - \kappa\mathcal{H}(p_h(s,a))$$

$$+ \hat{p}_h^t\overline{V}_{h+1}^t(s,a) - p_h\tilde{V}_{h+1}^t(s,a) + b_h^{p,t}(s,a).$$

By the definition of the event $\mathcal{E}^{\mathcal{H}}(\delta) \subseteq \mathcal{G}(\delta)$ we have

$$\kappa\mathcal{H}(\hat{p}_h^t(s,a)) + \kappa b_h^{\mathcal{H},t}(s,a) - \kappa\mathcal{H}(p_h(s,a)) \leq 2\kappa b_h^{\mathcal{H},t}(s,a),$$

for the next term we have

$$\hat{p}_h^t\overline{V}_{h+1}^t - p_h\tilde{V}_{h+1}^t(s,a) = \hat{p}_h^t[\overline{V}_{h+1}^t - \tilde{V}_{h+1}^t(s,a)] + [\hat{p}_h^t - p_h]V_{\lambda,h+1}^\star(s,a) + [p_h - \hat{p}_h^t][V_{\lambda,h+1}^\star - \tilde{V}_{h+1}^t](s,a).$$

By induction hypothesis

$$\hat{p}_h^t[\overline{V}_{h+1}^t - \tilde{V}_{h+1}^t(s,a)] \leq \hat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a).$$

Next we apply Lemma D.5 with $U = \overline{V}_{h+1}^t(s,a)$ and Theorem D.6

$$[\hat{p}_h^t - p_h]V_{\lambda,h+1}^\star(s,a) \leq b_h^{p,t}(s,a).$$

Finally, we apply Lemma D.4 and obtain

$$[p_h - \hat{p}_h^t][V_{\lambda,h+1}^\star - \tilde{V}_{h+1}^t](s,a) \leq \frac{1}{H}\hat{p}_h^t[V_{\lambda,h+1}^\star - \tilde{V}_{h+1}^t](s,a) + \frac{4H^2\text{R}_{\max} \cdot \beta^{\text{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Summing all these bounds up, we have

$$\overline{Q}_h^t(s,a) - \tilde{Q}_h^t(s,a) \leq \hat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{p,t}(s,a)$$

$$+ \frac{1}{H}\hat{p}_h^t[V_{\lambda,h+1}^\star - \tilde{V}_{h+1}^t](s,a) + \frac{4H^2\text{R}_{\max} \cdot \beta^{\text{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Notice that by Theorem D.6 and the induction hypothesis

$$\frac{1}{H}\hat{p}_h^t[V_{\lambda,h+1}^\star - \tilde{V}_{h+1}^t](s,a) \leq \frac{1}{H}\hat{p}_h^t[\overline{V}_{h+1}^t - \tilde{V}_{h+1}^t](s,a) \leq \frac{1}{H}\hat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a),$$

and by decomposing the transition bonus (14) to Bernstein bonus and correction term and applying Lemma D.8

$$b_h^{p,t}(s,a) = b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a) \le b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \tilde{V}_{h+1}^t](s,a)$$

$$\le b_h^{B,t}(s,a) + \frac{1}{H}\widehat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a),$$

thus

$$\overline{Q}_h^t(s,a) - \tilde{Q}_h^t(s,a) \le \left(1 + \frac{3}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a)$$

$$+ \frac{4H^2\mathrm{R}_{\max} \cdot \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} = G_h^t(s,a).$$

Second case.   In this case we have $\tilde{Q}_h^t(s,a) = \hat{r}_{\lambda,h}^t(s,a) + \widehat{p}_h^t\tilde{V}_{h+1}^t(s,a) - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a)$. Thus

$$\overline{Q}_h^t(s,a) - \tilde{Q}_h^t(s,a) \le 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a) + \widehat{p}_h^t[\overline{V}_{h+1}^t - \tilde{V}_{h+1}^t](s,a) + \frac{2}{H}\widehat{p}_h^t[\overline{V}_{h+1}^t - \underline{V}_{h+1}^t](s,a).$$

By Lemma D.8 and induction hypothesis we have

$$\overline{Q}_h^t(s,a) - \tilde{Q}_h^t(s,a) \le 2\kappa b_h^{\mathcal{H},t}(s,a) + 2b_h^{B,t}(s,a) + \left(1 + \frac{2}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a) \le G_h^t(s,a).$$

Conclusion.   From the two cases above we conclude

$$[\overline{Q}_h^t - \tilde{Q}_h^t](s,a) \le G_h^t(s,a).$$

Moreover, we have

$$\overline{V}_h^t(s) - \tilde{V}_h^t(s) = \pi_h^{t+1}\overline{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s))\pi_h^{t+1}\tilde{Q}_h^t(s) + \lambda\Phi(\pi_h^{t+1}(s)) = \pi_h^{t+1}[\overline{Q}_h^t - \tilde{Q}_h^t](s) \le \pi_h^{t+1}G_h^t(s).$$

The last inequality concludes the statement of Lemma D.7.   □

Lemma D.8. *Under the choice of Bernstein bonuses* (14)*, on event $\mathcal{G}(\delta)$ for any $t \in \mathbb{N}$ and any $(h,s,a) \in [H] \times \mathcal{S} \times \mathcal{A}$*

$$\tilde{Q}_h^t(s,a) \le \min\{Q_{\lambda,h}^{\pi^{t+1}}(s,a), \underline{Q}_h^t(s,a)\}, \qquad \tilde{V}_h^t(s) \le \min\{V_{\lambda,h}^{\pi^{t+1}}(s), \underline{V}_h^t(s)\}.$$

Proof. Proceed by backward induction over h. The case $h = H + 1$ is trivially true, assume that the statement holds for any $h' > h$ for a fixed h. Also let us fix t, s, a. By induction hypothesis we have

$$\tilde{Q}_h^t(s,a) \le r_{\kappa,h}(s,a) + p_h\tilde{V}_{h+1}^t \le r_{\kappa,h}(s,a) + p_h V_{\lambda,h+1}^{\pi^{t+1}}(s,a) = Q_{\lambda,h}^{\pi^{t+1}}(s,a).$$

In the same manner

$$\tilde{Q}_h^t(s,a) \le \hat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t\tilde{V}_{h+1}^t - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a)$$

$$\le \hat{r}_{\kappa,h}^t(s,a) + \widehat{p}_h^t\underline{V}_{h+1}^t - b_h^{p,t}(s,a) - \kappa b_h^{\mathcal{H},t}(s,a) = \underline{Q}_h^t(s,a).$$

Next, we prove the same inequalities for V-functions

$$\tilde{V}_h^t(s) = \pi_h^{t+1}(s)\tilde{Q}_h^t(s,a) - \lambda\Phi(\pi_h^{t+1}(s)) \le \pi_h^{t+1}(s)Q_{\lambda,h}^{\pi^{t+1}}(s,a) - \lambda\Phi(\pi_h^{t+1}(s)) = V_{\lambda,h}^{\pi^{t+1}}(s),$$

and

$$\tilde{V}_h^t(s) = \pi_h^{t+1}\tilde{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s)) \le \pi_h^{t+1}\underline{Q}_h^t(s) - \lambda\Phi(\pi_h^{t+1}(s)) \le \max_{\pi \in \Delta_{\mathcal{A}}}\left\{\pi\underline{Q}_h^t(s) - \lambda\Phi(\pi)\right\} = \underline{V}_h^t(s).$$

□

After defining a proper quantity for a stopping rule we may proceed with the final proof for sample-complexity of the presented algorithm `UCBVI-Ent`.

Theorem D.9. *Let $\delta \in (0,1)$. Then* `UCBVI-Ent` *algorithm with Bernstein bonuses* (14) *and a regularization-agnostic stopping rule τ is (ε, δ)-PAC for the best policy identification in regularized MDPs.*

Moreover, with probability at least $1 - \delta$ the stopping time τ is bounded as follows

$$\tau = \mathcal{O}\left(\frac{H^3 S A \mathrm{R}_{\max}^2 \log(SAH/\delta)L^4}{\varepsilon^2} + \frac{H^3 SA(\log(SAH/\delta) + SL) \cdot L}{\varepsilon}\right),$$

where $L = \mathcal{O}(\log(SAH\mathrm{R}_{\max}/\varepsilon)) + \log\log(SAH/\delta)$.

Proof. Notice that if $\tau = 0$, then our sample complexity bound is trivial, thus we assume that $\tau > 0$. Let us start from deriving an upper bound for $G_h^t(s,a)$ for $t < \tau, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$.

$$G_h^t(s,a) \leq 2b_h^{B,t}(s,a) + 2\kappa b_h^{\mathcal{H},t}(s,a) + \frac{4H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{3}{H}\right)\widehat{p}_h^t[\pi_{h+1}^{t+1}G_{h+1}^t](s,a)$$

$$\leq 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + \frac{23H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}$$

$$+ \left(1 + \frac{3}{H}\right)[\widehat{p}_h^t - p_h][\pi_{h+1}^{t+1}G_{h+1}^t](s,a) + \left(1 + \frac{3}{H}\right)p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a).$$

By Lemma D.4

$$[\widehat{p}_h^t - p_h][\pi_{h+1}^{t+1}G_{h+1}^t](s,a) \leq \frac{1}{H}p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a) + \frac{4H^2 \mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Also we have to replace the variance of the empirical model with the real variance of the value function for π^{t+1} in order to apply the law of total variance (Lemma C.3).

Apply Lemma H.9 and Lemma H.8

$$\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a) \leq 2\mathrm{Var}_{p_h}[\overline{V}_{h+1}^t](s,a) + \frac{4H^2 \mathrm{R}_{\max}^2 \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}$$

$$\leq 4\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) + 2H\mathrm{R}_{\max}p_h[\overline{V}_{h+1}^t - V_{\lambda,h+1}^{\pi^{t+1}}](s,a) + \frac{4H^2 \mathrm{R}_{\max}^2 \beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

In the proof of Lemma D.7 it was proven that

$$[\overline{V}_{h+1}^t - V_{\lambda,h+1}^{\pi^{t+1}}](s) \leq \pi_{h+1}^{t+1}G_{h+1}^t(s),$$

thus, combining with an inequality $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$

$$6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}[\overline{V}_{h+1}^t](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \leq 12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ 6\sqrt{p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a)\frac{2H\mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ \frac{12H\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}$$

To bound the second term, we use inequality $2\sqrt{ab} \leq a + b$

$$6\sqrt{p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a)\frac{2H\mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} \leq \frac{3}{H}p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a) + \frac{3H^2 \mathrm{R}_{\max}\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}.$$

Finally, we have the following bound on $G_h^t(s,a)$

$$G_h^t(s,a) \leq 12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$+ \frac{54H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{10}{H}\right)p_h[\pi_{h+1}^{t+1}G_{h+1}^t](s,a).$$

Notice that his inequality could be rewritten in the following form

$$G_h^t(s,a) \leq \mathbb{E}_{\pi^{t+1}}\left[12\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 2\kappa\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}\right.$$

$$\left.+ \frac{54H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s,a))}{n_h^t(s,a)} + \left(1 + \frac{10}{H}\right)G_{h+1}^t(s_{h+1}, a_{h+1})\Big|(s_h, a_h) = (s,a)\right],$$

thus by rolling out we have

$$\pi_1^{t+1}G_1^t(s_1) \leq \mathbb{E}_{\pi^{t+1}}\left[\underbrace{12\sum_{h=1}^{H}\left(1 + \frac{10}{H}\right)^h\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s_h, a_h)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s_h, a_h))}{n_h^t(s_h, a_h)}}}_{(\mathbf{A})}\right.$$

$$+ \underbrace{2\kappa\sum_{h=1}^{H}\left(1 + \frac{10}{H}\right)^h\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s_h, a_h))}{n_h^t(s_h, a_h)}}}_{(\mathbf{B})}$$

$$\left.+ \underbrace{\sum_{h=1}^{H}\left(1 + \frac{10}{H}\right)^h\frac{54H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta, n_h^t(s_h, a_h))}{n_h^t(s_h, a_h)}}_{(\mathbf{C})}\Big|s_1\right],$$

where $(1 + 10/H)^h \leq \mathrm{e}^{10}$ for any $h \in [H]$. Now we bound each term separately.

Term $(\mathbf{A})$. To bound this term, we apply Cauchy-Schwarz inequality

$$(\mathbf{A}) \leq 12\mathrm{e}^{10}\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\sqrt{\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}$$

$$\leq 12\mathrm{e}^{10}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a)} \cdot \sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

For the first multiplier we apply the law of total variance (Lemma C.3)

$$\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) \leq \sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\mathrm{Var}_{p_h}[V_{\lambda,h+1}^{\pi^{t+1}}](s,a) + \sum_{h,s}d_h^{\pi^{t+1}}(s)\mathrm{Var}_{\pi_h^{t+1}}[Q_{\lambda,h}^{\pi^{t+1}}](s)$$

$$= \sigma V_1^{\mathcal{H},\pi^{t+1}}(s_1) \leq H^2\mathrm{R}_{\max}^2.$$

Therefore,

$$(\mathbf{A}) \leq 24\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

Term (B). For this term we may apply Jensen's inequality

$$(\mathbf{B}) \le 2\kappa H \mathrm{e}^{10} \mathbb{E}_{\pi^{t+1}}\left[\frac{1}{H}\sum_{h=1}^{H}\sqrt{\frac{2\beta^{\mathcal{H}}(\delta, n_h^t(s_h, a_h))}{n_h^t(s_h, a_h)}}\,\Big|\,s_1\right] \le \kappa\sqrt{8H}\mathrm{e}^{10}\sqrt{\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta, n_h^t(s,a))}{n_h^t(s,a)}}.$$

By summing up and replacing counts by pseudo-counts by Lemma H.5 we obtain

$$\pi_1^{t+1}G_1^t(s_1) \le 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\kappa\mathrm{e}^{10}\sqrt{2H}\sqrt{\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}.$$

The last step is to notice that for $t < \tau$ we have $\pi_1^{t+1}G_1^t(s_1) \ge \varepsilon$, thus summing over all $t < \tau$ we have

$$(\tau-1)\varepsilon \le 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sum_{t=1}^{\tau-1}\sqrt{\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{conc}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\kappa\mathrm{e}^{10}\sqrt{2H}\sum_{t=1}^{\tau-1}\sqrt{\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathcal{H}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\sum_{t=1}^{\tau-1}\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{\beta^{\mathrm{KL}}(\delta, \overline{n}_h^t(s,a))}{\overline{n}_h^t(s,a)\vee 1}.$$

Also we notice that $\beta^{\mathrm{KL}}(\delta,\cdot), \beta^{\mathrm{conc}}(\delta,\cdot), \beta^{\mathcal{H}}(\delta,\cdot)$ are monotone, thus we may replace $\overline{n}_h^t(s,a)$ with a stopping time τ. Thus, by Jensen's inequality

$$(\tau-1)\varepsilon \le 48\mathrm{e}^{10}H\mathrm{R}_{\max}\sqrt{(\tau-1)\cdot\beta^{\mathrm{conc}}(\delta,\tau-1)}\sqrt{\sum_{t=1}^{\tau-1}\sum_{(h,s,a)\in[H]\times\mathcal{S}\times\mathcal{A}}d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\kappa\mathrm{e}^{10}\sqrt{2H\beta^{\mathcal{H}}(\delta,\tau)\cdot(\tau-1)}\sqrt{\sum_{t=1}^{\tau-1}\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}}$$

$$+ 4\mathrm{e}^{10}H^2\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta,\tau-1)\sum_{t=1}^{\tau-1}\sum_{h,s,a}d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1}.$$

Furthermore, notice

$$\sum_{t=1}^{\tau-1}d_h^{\pi^{t+1}}(s,a)\frac{1}{\overline{n}_h^t(s,a)\vee 1} = \sum_{t=1}^{\tau-1}\frac{\overline{n}_h^{t+1}(s,a)-\overline{n}_h^t(s,a)}{\overline{n}_h^t(s,a)\vee 1},$$

thus Lemma H.6 is applicable:

$$(\tau-1)\varepsilon \le 96\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{(\tau-1)H^3SA\log(\tau)\cdot\beta^{\mathrm{conc}}(\delta,\tau-1)}$$

$$+ 2\kappa\mathrm{e}^{10}\sqrt{2H^2SA\log^3(\tau)\beta^{\mathcal{H}}(\delta,\tau)\cdot(\tau-1)}$$

$$+ 8\mathrm{e}^{10}H^3SA\log(\tau)\mathrm{R}_{\max}\beta^{\mathrm{KL}}(\delta,\tau-1).$$

By the definitions of $\beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathcal{H}}$ we have the following inequality

$$(\tau - 1)\varepsilon \leq 96\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{(\tau - 1)H^3SA\log(\tau) \cdot (\log(16SAH/\delta) + 2\log(\mathrm{e}\tau))}$$

$$+ 12\kappa\mathrm{e}^{10}\sqrt{H^2SA(\tau - 1)\log^3(\tau) \cdot (\log(4SAH/\delta) + 2\log(\mathrm{e}\tau))}$$

$$+ 16\mathrm{e}^{10}H^3SA\log(\tau)\mathrm{R}_{\max}(\log(4SAH/\delta) + S\log(\mathrm{e}\tau)).$$

Under assumption $\tau \geq 2$ we can proceed with the further simplifications

$$\tau\varepsilon \leq 216\mathrm{e}^{10}\mathrm{R}_{\max}\sqrt{\tau H^3SA\log^3(\tau) \cdot (\log(16SAH/\delta) + 2\log(\mathrm{e}\tau))}$$

$$+ 32\mathrm{e}^{10}H^3SA\log(\tau)\mathrm{R}_{\max}(\log(16SAH/\delta) + S\log(\mathrm{e}\tau)). \tag{18}$$

Let us define the following constants

$$A = 216\mathrm{e}^{10}\mathrm{R}_{\max} \cdot \sqrt{\frac{H^3SA}{\varepsilon^2}}, \quad B = \log(16SAH/\delta), \quad C = \frac{32\mathrm{e}^{10} \cdot H^3SA\mathrm{R}_{\max}}{\varepsilon}.$$

Then inequality (18) has the following form

$$\tau \leq A\sqrt{\tau(B + 2\log(\mathrm{e}\tau)) \cdot \log^3(\tau)} + C(B + S\log(\mathrm{e}\tau))\log\tau.$$

First, we obtain a loose inequality on τ. Let us use the inequality $\log(x) \leq x^\beta/\beta$ for any $x > 0, \beta > 0$ with different β for each logarithm

$$\tau \leq A\sqrt{216\tau(B + 4(\mathrm{e}\tau)^{1/4})\tau^{1/2}} + 4C(B + 8S/3(\mathrm{e}\tau)^{3/8})\tau^{1/4}$$

$$\Rightarrow \tau^{3/4} \leq \tau^{3/8}\left(6A\sqrt{6(B + 4\mathrm{e}^{1/4})} + 12CSe^{3/8}\right) + 4CB.$$

Notice that the solution to the inequality $x^2 \leq ax + b$ could be upper-bounded as follows

$$x \leq \frac{a + \sqrt{a^2 + 4b}}{2} \leq a + \sqrt{b},$$

thus

$$\tau^{3/8} \leq \left(6A\sqrt{6(B + 4\mathrm{e}^{1/4})} + 12CSe^{3/8}\right) + 2\sqrt{CB}.$$

Define $L = 8/3\log\left(6A\sqrt{6(B + 4\mathrm{e}^{1/4})} + 12CSe^{3/8} + 2\sqrt{CB}\right) = \mathcal{O}(\log(SAH\mathrm{R}_{\max}/\varepsilon) + \log\log(SAH/\delta))$ and we have $\log(\tau) \leq L$. Then we have that the solution to (18) is a subset of solutions to

$$\tau \leq A\sqrt{\tau(B + 2(1 + L)) \cdot L^3} + C(B + S(1 + L))L,$$

solving this inequality we obtain the bound

$$\tau \leq 2A^2(B + 2(1 + L))L^3 + 2CB(S(1 + L))L.$$

$\square$

After this general result we state the bound for the MTEE problem that was stated in the main text.

Theorem D.10. *For all $\varepsilon > 0$ and $\delta \in (0, 1)$ the* `UCBVI-Ent` *algorithm is (ε, δ)-PAC for MTEE. Moreover, with probability at least $1 - \delta$*

$$\tau \leq \mathcal{O}\left(\frac{H^3SA\log^2(SA)\log(SAH/\delta) \cdot L^4}{\varepsilon^2} + \frac{H^3SA(\log(SAH/\delta) + SL) \cdot L}{\varepsilon}\right),$$

where $L = \log(SAH/\varepsilon) + \log\log(SAH/\delta)$.

Proof. Fix $\Phi(\pi) = -\mathcal{H}(\pi), \kappa = \lambda = 1$ and $r_{\max} = 0$. Since $\mathcal{H}(\pi)$ is 1-strongly convex with respect to ℓ_1-norm, we have $r_A = 1$. Also we automatically have $\mathrm{R}_{\max} = \log(SA)$. In this setting, Theorem D.9 yields the desired statement. $\square$

E. Fast Rates for MTEE and Regularized MDPs

In this section we describe an algorithm that will achieve $\widetilde{\mathcal{O}}(\text{poly}(S, A, H)/\varepsilon)$ sample complexity for regularized MDPs. Additionally, we show that this algorithm could be used for reward-free exploration under regularization.

E.1. `RL-Explore-Ent` Algorithm

We leverage the reward-free exploration approach by Jin et al. (2020). Our algorithm is split into two phases: the first phase is devoted to reward-free exploration, and on the second phase the collected samples are used to build estimates of transition probabilities and entropy of transitions. The main idea is that regularization allows us to collect much smaller number of samples to control the policy error.

Exploration phase We first learn a policy that visit uniformly the MDP. To this aim for each state $s' \in \mathcal{S}$ at each step $h' \in [H]$ we build the reward that put one on this state at step h and zeros everywhere else $r_h(s, a) = \mathbb{1}\{(s, h) = (s', h')\}$. We note that the reward function does not depend on action taken. We then run the `EULER` algorithm for N_0 episodes in the MDP equipped with the reward r and denote by $\tilde{\Pi}_{s',h'}$ the set of N_0 policies used by `EULER` to interact with the MDP. We modify this set of policies by forcing to act uniformly at the goal state s' into the set

$$
\Pi_{s',h'} = \left\{ \pi'_h(a|s) = \begin{cases} 1/A & \text{if } s = s', h = h' \\ \pi_h(s, a) & \text{else} \end{cases} : \pi \in \tilde{\Pi}_{s',h'} \right\}.
$$

We define the (non-Markovian) policy π^{mix} as the uniform mixture of the policies $\{\pi \in \Pi_{s,h}, (s, h) \in \mathcal{S} \times [H]\}$ we just constructed. As proved by Jin et al. (2020) the policy π^{mix} is built such that it will visit almost uniformly all the states that can be reached in the MDP from the initial state. Before precising this property we need to introduce the notion of significant state.

Definition E.1. A state s at step h is called ε'-significant if there exists a policy π such that the visitation probability of s under policy π is greater than ε':

$$
\max_\pi d_h^\pi(s) \geq \varepsilon'.
$$

The set of all ε'-significant state-step pairs is called $S_{\varepsilon'}$.

We reproduce here the result by Jin et al. (2020) that shows that the policy π^{mix} will visit any significant state with a large enough probability.

Theorem E.2 (Theorem 3.3 by Jin et al. 2020). *There exists an absolute constant $c > 0$ such that for any $\varepsilon' > 0$ and $\delta \in (0, 1)$, if we set the parameter $N_0 \geq cS^2AH^4L/\varepsilon'$ where $L = \log(SAH/(\delta\varepsilon'))$, then with probability at least $1 - \delta/3$ the following event holds*

$$
\mathcal{E}^{\text{RF-RL-Explore}}(\delta, \varepsilon') = \left\{ \forall (s, h) \in S_{\varepsilon'}, \forall a \in \mathcal{A}, \forall \pi : \frac{d_h^\pi(s, a)}{\mu_h(s, a)} \leq 2SAH \right\},
$$

where we denote the visitation distribution of policy π^{mix} by $\mu_h(s, a) = d_h^{\pi^{\text{mix}}}(s, a)$.

Remark E.3. Note that the space complexity of `RL-Explore-Ent` is very large since we need to store all the intermediate policies in order to construct π^{mix}.

This policy π^{mix} is then used to collect N new independent trajectories $(z_n)_{n \in [N]}$ where $z_n = (s_1^n, a_1^n, \ldots, s_H^n, a_H^n, s_{H+1}^n)$ by following the policy π^{mix} in the original MDP. Next define the set $\mathcal{D} = \{(s_h^n, a_h^n, s_{h+1}^n), h \in [H], n \in [N]\}$ consisting of the transitions in the sampled trajectories.

Planning phase Given the transitions collected in the exploration phase we estimate the transition probability distributions and then plan in the estimated MDP with the Bellman equations for MTEE to obtain a maximum trajectory entropy policy.

Using the dataset $\mathcal{D}$, we construct estimates of transition probabilities $\{\widehat{p}_h\}_{h \in [H]}$. We first define the number of visits of a state action pair (s, a) at step h and the number of transitions for (s, a) at step h to a states s' observed in the dataset $\mathcal{D}$,

$$
n_h(s, a) = \sum_{n=1}^N \mathbb{1}\{(s_h^n, a_h^n) = (s, a)\} \quad n_h(s'|s, a) = \sum_{n=1}^N \mathbb{1}\{(s_h^n, a_h^n, s_{h+1}^n) = (s, a, s')\},
$$

The transitions are estimated using the maximum likelihood method:

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0 \\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}. \tag{19}$$

Given these estimates, we can define an empirical version of the regularized Bellman equations for MTEE

$$\begin{aligned} \widehat{Q}_{\lambda,h}^\pi(s,a) &= r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h(s,a)) + \widehat{p}_h\widehat{V}_{\lambda,h+1}^\pi(s,a) \\ \widehat{V}_{\lambda,h}^\pi(s) &= \pi\widehat{Q}_{\lambda,h}^\pi(s) - \lambda\Phi(\pi). \end{aligned} \tag{20}$$

Then the output policy $\widehat{\pi}$ is the solution to the optimal regularized Bellman equations

$$\begin{aligned} \widehat{Q}_{\lambda,h}^\star(s,a) &= r_h(s,a) + \kappa\mathcal{H}(\widehat{p}_h(s,a)) + \widehat{p}_h\widehat{V}_{\lambda,h+1}^\star(s,a) \\ \widehat{V}_{\lambda,h}^\star(s) &= \max_\pi\left\{\pi\widehat{Q}_{\lambda,h}^\star(s) - \lambda\Phi(\pi)\right\} \\ \widehat{\pi}_h(s) &= \arg\max_\pi\left\{\pi\widehat{Q}_{\lambda,h}^\star(s) - \lambda\Phi(\pi)\right\}. \end{aligned} \tag{21}$$

We call this algorithm `RL-Explore-Ent`. Notably, we can extend this algorithm to the setting of the changing rewards by solving (21) with new reward functions $r_h(s,a)$. The detailed description of the algorithm is presented in Algorithm 2. The only difference between our algorithm and `RF-RL-Explore` by Jin et al. (2020) is the use of a smaller number of trajectories N and solving regularized Bellman equations instead of usual one.

E.2. Concentration Events

In this section we describe all required concentration events.

Let $\beta^{\mathrm{conc}}\colon (0,1) \times \mathbb{N} \to \mathbb{R}_+$ and $\beta^{\mathrm{cnt}}\colon (0,1) \to \mathbb{R}_+$ be some functions defined later on in Lemma E.4. We define the following favorable events

$$\mathcal{E}^{\mathrm{conc}}(N,\delta) \triangleq \left\{ \forall h \in [H], \forall G\colon \mathcal{S} \to [0, H\mathrm{R}_{\max}], \forall \nu\colon \mathcal{S} \to \mathcal{A} : \right.$$

$$\left. \mathbb{E}_{(s,a)\sim\mu_h}\left[([\widehat{p}_{h'} - p_{h'}]G(s,a))^2\mathbb{1}\{\nu(s) = a\}\right] \leq \frac{CH^2\mathrm{R}_{\max}^2 S \cdot \beta^{\mathrm{conc}}(\delta,N)}{N} \right\},$$

$$\mathcal{E}^{\mathcal{H}}(N,\delta) \triangleq \left\{ \forall h \in [H] : \mathbb{E}_{(s,a)\sim\mu_h}\left[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\right] \leq \frac{12S^2A\log^2(SN) \cdot \beta^{\mathcal{H}}(\delta)}{N} \right\},$$

where C is a some absolute constant. We also introduce two intersections of these events of interest and $\mathcal{E}^{\mathrm{RF-RL-Explore}}(\delta)$, defined in Theorem E.2: $\mathcal{G}(N,\delta,\varepsilon') \triangleq \mathcal{E}^{\mathrm{conc}}(N,\delta) \cap \mathcal{E}^{\mathcal{H}}(\delta) \cap \mathcal{E}^{\mathrm{RF-RL-Explore}}(\delta,\varepsilon')$. We prove that for the right choice of the functions $\beta^{\mathrm{conc}}, \beta^{\mathcal{H}}$ the above events hold with high probability.

Lemma E.4. *For any $\delta \in (0,1), \varepsilon' > 0, N \in \mathbb{N}$ and for the following choices of functions β,*

$$\beta^{\mathrm{conc}}(\delta,N) \triangleq \log(3AH\mathrm{R}_{\max}N/\delta),$$
$$\beta^{\mathcal{H}}(\delta) \triangleq \log(12SAH/\delta),$$

it holds that

$$\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geq 1 - \delta/3, \qquad \mathbb{P}[\mathcal{E}^{\mathcal{H}}(\delta)] \geq 1 - \delta/3.$$

In particular, $\mathbb{P}[\mathcal{G}(\delta,N,\varepsilon')] \geq 1 - \delta$.

Proof. Holds from an application of Lemma G.6, Lemma G.5, Theorem E.2 and union bound. $\square$

E.3. Sample Complexity Proof

In this section we provide the sample complexity result of `RL-Explore-Ent` algorithm in the simple BPI setting and in the reward free setting.

Theorem E.5. *Algorithm* `RL-Explore-Ent` *with parameters* $N_0 = \Omega\left(\frac{H^7 S^3 A r_A^2 \cdot R_{\max}^2 \cdot L}{\varepsilon\lambda}\right)$ *and* $N \geq \Omega\left(\frac{H^6 S^3 A r_A^2 R_{\max}^2 L^3}{\varepsilon\lambda}\right)$ *is* (ε, δ)-*PAC for the best policy identification in regularized MDPs, where* $L = \log(SAH/(\varepsilon\lambda\delta))$. *The sample complexity is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8 S^4 A r_A^2 R_{\max}^2}{\varepsilon\lambda}\right).$$

Proof. Let us start from exploiting the strong convexity of the regularizer. This property is given by Lemma E.8

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\widehat{\pi}}(s_1) \leq \frac{r_A^2}{2\lambda} \sum_{h=1}^{H} \mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}} - Q_{\lambda,h}^\star\right)^2(s_h, a)\middle|s_1\right].$$

Next we study each separate term in this decomposition. By the definition of $\widehat{\pi}$ and $\pi^\star$ we have

$$\widehat{Q}_{\lambda,h}^{\pi^\star}(s,a) - Q_{\lambda,h}^{\pi^\star}(s,a) \leq \widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^\star(s,a) \leq \widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^{\widehat{\pi}}(s,a),$$

thus

$$\left(\widehat{Q}_{\lambda,h}^{\pi^\star}(s,a) - Q_{\lambda,h}^{\pi^\star}(s,a)\right)^2 \leq \max\left\{\left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^\star(s,a)\right)^2, \left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^{\widehat{\pi}}(s,a)\right)^2\right\},$$

and by an inequality $\max\{a, b\} \leq a + b$ for positive a, b we have

$$\left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^\star(s,a)\right)^2 \leq \left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s,a) - Q_{\lambda,h}^{\widehat{\pi}}(s,a)\right)^2 + \left(\widehat{Q}_{\lambda,h}^{\pi^\star}(s,a) - Q_{\lambda,h}^{\pi^\star}(s,a)\right)^2.$$

Therefore, the policy error decomposes as follows

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\widehat{\pi}}(s_1) \leq \frac{r_A^2}{2\lambda} \sum_{h=1}^{H}\left(\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\widehat{\pi}}(s_h, a) - Q_{\lambda,h}^{\widehat{\pi}}(s_h, a)\right)^2 \mid s_1\right]\right.$$
$$\left. + \mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi^\star}(s_h, a) - Q_{\lambda,h}^{\pi^\star}(s_h, a)\right)^2 \mid s_1\right]\right).$$

Next we assume that the event $\mathcal{G}(N, \delta, \varepsilon')$ holds for the values N and ε' that will be specified later. Then Lemma E.9 applied $2H$ times yields

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\widehat{\pi}}(s_1) \leq \frac{r_A^2}{\lambda}\left(\frac{48 S^3 H^4 A R_{\max}^2 \log^2(N) \log(12SAH/\delta)}{N}\right.$$
$$\left. + \frac{4C H^6 R_{\max}^2 S^2 A \cdot (\log(3AHR_{\max}/\delta) + \log(N))}{N} + 2SH^3 R_{\max}^2 \cdot \varepsilon'\right).$$

Next we take

$$\varepsilon' = \frac{\lambda\varepsilon}{4r_A^2 SH^3 R_{\max}^2},$$

that requires to take $N_0 \geq \frac{cH^7 S^3 A r_A^2 \cdot R_{\max}^2 \cdot L}{\varepsilon\lambda}$ for an absolute constant $c > 0$ and $L = \log(SAH/\delta) + \log(1/(\varepsilon\lambda))$. This yields $\widetilde{\mathcal{O}}(H^8 S^4 A r_A^2 R_{\max}^2/(\varepsilon\lambda))$ sample complexity of the first phase, since we need N_0 samples for each $(s, h) \in \mathcal{S} \times [H]$. Under this choice, we have

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^{\widehat{\pi}}(s_1) \leq \varepsilon/2 + \frac{r_A^2}{\lambda N} \cdot \left((48 + 4C)S^3 AH^6 R_{\max}^2 \log^2(N) \cdot \log(12SAHR_{\max}/\delta)\right).$$

To make the second part smaller than ε, we have to analyze the following inequality

$$\log^2(N)/N \cdot B \leq \varepsilon$$

and upper bound its minimal solution that we will call $N^\star$. To do it, we first use a simple numeric bound $\log(N) \leq 4N^{1/4}$ and obtain a simple estimate $N \geq 256B^2/\varepsilon^2$, thus the minimal solution $N^\star \leq 256B^2/\varepsilon^2$. Therefore, we can assume that $\log(N^\star) \leq 2\log(16B/\varepsilon) = \mathcal{O}(\log(SAH\mathrm{R}_{\max}/\varepsilon + \log(1/\delta))$, thus taking

$$N \geq N^\star = \Omega\left(\frac{H^6 S^3 A r_A^2 \mathrm{R}_{\max}^2 \log^2(SAH\mathrm{R}_{\max}/\varepsilon + \log(1/\delta)) \cdot \log(SAH\mathrm{R}_{\max}/\delta)}{\varepsilon\lambda}\right).$$

is enough to guarantee that the policy error is smaller than ε. $\qquad\square$

Notice that in the proof we do not rely on one particular reward function, since the only we need is conditioning on event $\mathcal{G}(\delta)$ that does not depend on the particular reward function.

Corollary E.6. *Algorithm* `RL-Explore-Ent` *for a choice* $N_0 = \Omega\left(\frac{H^7 S^3 A r_A^2 \cdot \mathrm{R}_{\max}^2 \cdot L}{\varepsilon\lambda}\right)$ *and* $N \geq \Omega\left(\frac{H^6 S^3 A r_A^2 \mathrm{R}_{\max}^2 L^3}{\varepsilon\lambda}\right)$ *outputs* ε*-optimal policies for an arbitrary number of reward functions in regularized MDPs. The sample complexity is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8 S^4 A r_A^2 \mathrm{R}_{\max}^2}{\varepsilon\lambda}\right).$$

Finally, we provide a formal proof for application of this algorithm to the MTEE problem, that is a simple application of the results above.

Theorem E.7. *Algorithm* `RL-Explore-Ent` *with parameters* $N_0 = \Omega\left(\frac{H^7 S^3 A \cdot L^3}{\varepsilon}\right)$ *and* $N = \Omega\left(\frac{H^6 S^3 A L^5}{\varepsilon}\right)$ *is* (ε, δ)*-PAC for the MTEE problem, where* $L = \log(SAH/(\varepsilon\delta))$. *The total sample complexity* $SHN_0 + N$ *is bounded by*

$$\widetilde{\mathcal{O}}\left(\frac{H^8 S^4 A}{\varepsilon}\right).$$

Proof. Fix $\Phi(\pi) = -\mathcal{H}(\pi), \kappa = \lambda = 1$ and $r_{\max} = 0$. By 1-strong convexity of $-\mathcal{H}(\pi)$ with respect to ℓ_1-norm, its dual is 1-strongly convex with respect to ℓ_∞ norm, yielding $r_A = 1$. Also we have $\mathrm{R}_{\max} = \log(SA)$, thus by Theorem E.5 we conclude the statement. $\qquad\square$

E.4. Technical Lemmas

Lemma E.8. *Let* π *be a greedy policy with respect to regularized Q-values* $\overline{Q}_{\lambda,h}(s,a) : \pi_h(s) = \arg\max_\pi \{\pi\overline{Q}_{\lambda,h}(s) - \lambda\Phi(\pi)\}$. *Then the following error decomposition holds*

$$V_{\lambda,1}^\star(s_1) - V_{\lambda,1}^\pi(s_1) \leq \frac{r_A^2}{2\lambda}\mathbb{E}_\pi\left[\sum_{h=1}^H \max_{a\in\mathcal{A}}(\overline{Q}_{\lambda,h} - Q_{\lambda,h}^\star)^2(s_h,a) \mid s_1\right],$$

where r_A *is a constant defined in* (11).

Proof. First, we formulate the statement dependent of h

$$V_{\lambda,h}^\star(s) - V_{\lambda,h}^\pi(s) \leq \frac{r_A^2}{2\lambda}\mathbb{E}_\pi\left[\sum_{h'=h}^H \max_{a\in\mathcal{A}}(\overline{Q}_{\lambda,h'} - Q_{\lambda,h'}^\star)^2(s_{h'},a) \mid s_h = s\right]. \tag{22}$$

Notice that for $h = 1$ and $s = s_1$ the initial statement is recovered. We proceed by induction over h. The initial case $h = H+1$ is trivial, next we assume that the statement (22) is true for any $h' > h$.

We start the analysis from understanding the policy error by applying the smoothness of F_λ for any h.

$$\begin{aligned}
V_{\lambda,h}^\star(s) - V_{\lambda,h}^\pi(s) &= F_\lambda(Q_{\lambda,h}^\star(s,\cdot)) - \left(\pi_h Q_{\lambda,h}^\pi(s,\cdot) - \lambda\Phi(\pi_h(s))\right) \\
&\leq F_\lambda(\overline{Q}_h)(s) + \langle\nabla F_\lambda(\overline{Q}_h(s,\cdot)), Q_{\lambda,h}^\star(s,\cdot) - \overline{Q}_h(s,\cdot)\rangle + \frac{1}{2\lambda}\|\overline{Q}_h - Q_{\lambda,h}^\star\|_*^2(s) \\
&\quad - \left(\pi_h Q_{\lambda,h}^\pi(s,\cdot) - \lambda\Phi(\pi_h(s))\right).
\end{aligned}$$

Next we recall that

$$\pi_h(s) = \nabla F(\overline{Q}_h(s, \cdot)), \quad F(\overline{Q}_h)(s) = \pi_h \overline{Q}_h(s) - \lambda \Phi(\pi_h(s)),$$

thus we have

$$F(\overline{Q}_h)(s) - \left(\pi_h Q_{\lambda,h}^\pi(s, \cdot) - \lambda \Phi(\pi_h(s)) \right) = \pi_h[\overline{Q}_h - Q_{\lambda,h}^\pi](s)$$

and, by Bellman equations

$$V_{\lambda,h}^\star(s) - V_{\lambda,h}^\pi(s) \le \pi_h \left[Q_{\lambda,h}^\star - Q_{\lambda,h}^\pi \right](s) + \frac{1}{2\lambda} \|\overline{Q}_h - Q_{\lambda,h}^\star\|_*^2(s)$$

$$\le \pi_h p_h \left[V_{\lambda,h+1}^\star - V_{\lambda,h+1}^\pi \right](s) + \frac{1}{2\lambda} \|\overline{Q}_h - Q_{\lambda,h}^\star\|_*^2(s).$$

Applying norm equivalence (11) we have

$$V_{\lambda,h}^\star(s) - V_{\lambda,h}^\pi(s) \le \mathbb{E}_\pi \left[\frac{r_A^2}{2\lambda} \|\overline{Q}_h - Q_{\lambda,h}^\star\|_\infty^2(s_h) + V_{\lambda,h+1}^\star(s_{h+1}) - V_{\lambda,h+1}^\pi(s_{h+1}) \mid s_h = s \right].$$

By induction hypothesis we conclude the statement. $\qquad \square$

Lemma E.9. *For any policy π the following holds on event $\mathcal{G}(N, \delta, \varepsilon')$ defined in Lemma E.4 for any $h \in [H]$*

$$\mathbb{E}_{\widehat{\pi}} \left[\max_{a \in \mathcal{A}} \left(\widehat{Q}_{\lambda,h}^\pi - Q_{\lambda,h}^\pi \right)^2(s_h, a) \mid s_1 \right] \le \frac{48 S^3 H^3 A \mathrm{R}_{\max}^2 \log^2(N) \beta^{\mathcal{H}}(\delta)}{N} + \frac{4 C H^5 \mathrm{R}_{\max}^2 S^2 A \cdot \beta^{\mathrm{conc}}(\delta, N)}{N}$$
$$+ 2 S H^2 \mathrm{R}_{\max}^2 \cdot \varepsilon'.$$

Proof. By performance-difference Lemma C.4 and form of the rewards stated in (20) we have for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$

$$\widehat{Q}_{\lambda,h}^\pi(s, a) - Q_{\lambda,h}^\pi(s, a) = \kappa \mathbb{E}_\pi \left[\sum_{h'=h}^H \mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})) \mid (s_h, a_h) = (s, a) \right]$$
$$+ \mathbb{E}_\pi \left[\sum_{h'=h}^H [\widehat{p}_{h'} - p_{h'}] \widehat{V}_{\lambda,h'+1}^\pi(s_{h'}, a_{h'}) \mid (s_h, a_h) = (s, a) \right].$$

Next we analyze all required expectation for one fixed value $h \in [H]$. By Jensen's inequality and a simple algebraic inequality $(a + b)^2 \le 2a^2 + 2b^2$

$$\left(\widehat{Q}_{\lambda,h}^\pi(s, a) - Q_{\lambda,h}^\pi(s, a) \right)^2 \le 2\kappa^2 \mathbb{E}_\pi \left[\left(\sum_{h'=h}^H \mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})) \right)^2 \bigg| (s_h, a_h) = (s, a) \right]$$
$$+ 2 \mathbb{E}_\pi \left[\left(\sum_{h'=h}^H [\widehat{p}_{h'} - p_{h'}] \widehat{V}_{\lambda,h'+1}^\pi(s_{h'}, a_{h'}) \right)^2 \bigg| (s_h, a_h) = (s, a) \right].$$

By Cauchy–Schwarz inequality we have the final form

$$\left(\widehat{Q}_{\lambda,h}^\pi(s, a) - Q_{\lambda,h}^\pi(s, a) \right)^2 \le 2H \mathbb{E}_\pi \left[\sum_{h'=h}^H \kappa^2 (\mathcal{H}(\widehat{p}_{h'}(s_{h'}, a_{h'})) - \mathcal{H}(p_{h'}(s_{h'}, a_{h'})))^2 \right.$$
$$\left. + \sum_{h'=h}^H \left([\widehat{p}_{h'} - p_{h'}] \widehat{V}_{\lambda,h'+1}^\pi \right)^2 (s_{h'}, a_{h'}) \mid (s_h, a_h) = (s, a) \right] \tag{23}$$

To connect this conditional expectation in the right-hand side of (23) with conditional expectation over $\widehat{\pi}$, we define the following policy

$$\widetilde{\pi}_{h'}(a|s) = \begin{cases} \widehat{\pi}_{h'}(a|s) & h' < h \\ \mathbb{1}\{a = \arg\max_{a \in \mathcal{A}} \left\{ \left(\widehat{Q}_h^\pi(s, a) - Q_h^\pi(s, a) \right)^2 \right\} & h' = h \\ \pi_h(a|s) & h' > h. \end{cases}$$

Since we do not change the policy π for steps greater than h, we can replace π with $\widetilde{\pi}$ in (23). Additionally, since this policy is equal to $\widehat{\pi}$ for the first $h-1$ steps, the distribution $d_h^{\widehat{\pi}}(s_h)$ is equal to $d_h^{\widetilde{\pi}}(s_h)$:

$$d_h^{\widehat{\pi}}(s_h) = \sum_{s_1,a_1,\ldots,s_{h-1},a_{h-1}} \widehat{\pi}_1(a_1|s_1)\left(\prod_{h'=1}^{h-1} p_{h'-1}(s_{h'}|s_{h'-1},a_{h'-1})\widehat{\pi}_{h'}(a_{h'}|s_{h'})\right) \cdot p_{h-1}(s_h|s_{h-1},a_{h-1})$$

$$= \sum_{s_1,a_1,\ldots,s_{h-1},a_{h-1}} \widetilde{\pi}_1(a_1|s_1)\left(\prod_{h'=1}^{h-1} p_{h'-1}(s_{h'}|s_{h'-1},a_{h'-1})\widetilde{\pi}_{h'}(a_{h'}|s_{h'})\right) \cdot p_{h-1}(s_h|s_{h-1},a_{h-1}) = d_h^{\widetilde{\pi}}(s_h).$$

Therefore

$$\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s_h,a)\mid s_1\right] = \sum_s d_h^{\widehat{\pi}}(s)\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s,a)$$

$$= \sum_s d_h^{\widetilde{\pi}}(s)\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s,a)$$

$$= \mathbb{E}_{\widetilde{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s_h,a)\mid s_1\right] = \mathbb{E}_{\widetilde{\pi}}\left[\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s_h,a_h)\mid s_1\right].$$

Next we show that in (23) we can make a change of policy from π to $\widetilde{\pi}$. It is enough to show that the required marginal distributions are equal for all $h' \geq h$, i.e. for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ it holds $\mathbb{P}_\pi[(s_{h'},a_{h'})|(s_h,a_h)] = \mathbb{P}_{\widetilde{\pi}}[(s_{h'},a_{h'})|(s_h,a_h)]$. For $h' = h$ this probability is an indicator on (s_h,a_h), so it does not depend on policy. For the general case $h' > h$ we can use Markov property and imply

$$\mathbb{P}_\pi[(s_{h'},a_{h'})|(s_h,a_h)] = \sum_{(s_{h+1},a_{h+1},\ldots,s_{h'-1},a_{h'-1})} \prod_{i=h}^{h'-1} \pi_{i+1}(a_{i+1}|s_{i+1})p_i(s_{i+1}|s_i,a_i)$$

$$= \sum_{(s_{h+1},a_{h+1},\ldots,s_{h'-1},a_{h'-1})} \prod_{i=h}^{h'-1} \widetilde{\pi}_{i+1}(a_{i+1}|s_{i+1})p_i(s_{i+1}|s_i,a_i) = \mathbb{P}_{\widetilde{\pi}}[(s_{h'},a_{h'})|(s_h,a_h)].$$

Therefore, we can make change of measure in (23) and obtain

$$\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s_h,a)\mid s_1\right] \leq 2H\mathbb{E}_{\widetilde{\pi}}\left[\mathbb{E}_{\widetilde{\pi}}\left[\sum_{h'=h}^{H}\kappa^2(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'})) - \mathcal{H}(p_{h'}(s_{h'},a_{h'})))^2\mid s_h,a_h\right]\right.$$

$$\left.+ \mathbb{E}_{\widetilde{\pi}}\left[\sum_{h'=h}^{H}\left([\widehat{p}_{h'} - p_{h'}]\widehat{V}_{\lambda,h'+1}^{\pi}\right)^2(s_{h'},a_{h'})\mid s_h,a_h\right]\bigg| s_1\right].$$

By the properties of conditional expectation we can eliminate the inner expectation and obtain the following upper bound

$$\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda,h}^{\pi} - Q_{\lambda,h}^{\pi}\right)^2(s_h,a)\mid s_1\right] \leq 2H\kappa^2\sum_{h'=h}^{H}\mathbb{E}_{\widetilde{\pi}}\left[(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'})) - \mathcal{H}(p_{h'}(s_{h'},a_{h'})))^2\mid s_1\right]$$

$$+ 2H\sum_{h'=h}^{H}\mathbb{E}_{\widetilde{\pi}}\left[\left([\widehat{p}_{h'} - p_{h'}]\widehat{V}_{\lambda,h'+1}^{\pi}\right)^2(s_{h'},a_{h'})\bigg| s_1\right].$$

Applying Lemma E.10 and the definition of $\mathcal{E}^{\mathcal{H}}(\delta)$ we have

$$\mathbb{E}_{\widetilde{\pi}}\left[(\mathcal{H}(\widehat{p}_{h'}(s_{h'},a_{h'})) - \mathcal{H}(p_{h'}(s_{h'},a_{h'})))^2\mid s_1\right] \leq 2SAH\mathbb{E}_{(s,a)\sim\mu_{h'}}\left[(\mathcal{H}(\widehat{p}_{h'}(s,a)) - \mathcal{H}(p_{h'}(s,a)))^2\right] + S\log^2(S)\varepsilon'$$

$$\leq \frac{24S^3HA\log^2(SN)\beta^{\mathcal{H}}(\delta)}{N} + S\log^2(S)\varepsilon'.$$

In the same way by Lemma E.10 and the definition of event $\mathcal{E}^{\mathrm{conc}}(N, \delta)$

$$\mathbb{E}_{\widetilde{\pi}}\left[\left([\widehat{p}_{h'} - p_{h'}]\widehat{V}_{\lambda, h'+1}^{\pi}\right)^2(s_{h'}, a_{h'}) \mid s_1\right] \leq 2SAH\mathbb{E}_{(s,a)\sim\mu_{h'}}\left[\left([\widehat{p}_{h'} - p_{h'}]\widehat{V}_{\lambda, h'+1}^{\pi}\right)^2(s, a)\right] + SH^2\mathrm{R}_{\max}^2\varepsilon'$$

$$\leq \frac{2CH^3\mathrm{R}_{\max}^2 S^2 A \cdot \beta^{\mathrm{conc}}(\delta, N)}{N} + SH^2\mathrm{R}_{\max}^2\varepsilon'.$$

Combining these two upper bounds, we have

$$\mathbb{E}_{\widehat{\pi}}\left[\max_{a\in\mathcal{A}}\left(\widehat{Q}_{\lambda, h}^{\pi} - Q_{\lambda, h}^{\pi}\right)^2(s_h, a) \mid s_1\right] \leq \frac{48S^3H^3A\kappa^2\log^2(SN)\beta^{\mathcal{H}}(\delta)}{N} + \frac{4CH^5\mathrm{R}_{\max}^2 S^2 A \cdot \beta^{\mathrm{conc}}(\delta, N)}{N}$$

$$+ S(H^2\mathrm{R}_{\max}^2 + \kappa^2\log^2(S))\varepsilon'.$$

Since $\kappa^2\log^2(s) \leq \mathrm{R}_{\max}^2$ and $H \geq 1$, we conclude the statement. $\qquad\square$

Lemma E.10. *For any bounded function $f \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+, f(s, a) \leq B$ for any policy π and step h on event $\mathcal{E}^{RF\text{-}RL\text{-}Explore}(\delta, \varepsilon')$ the following holds*

$$\mathbb{E}_{\pi}[f(s_h, a_h)|s_1] \leq 2SAH\mathbb{E}_{(s,a)\sim\mu_h}[f(s, a)] + BS\varepsilon'.$$

Proof. Recall $S_{\varepsilon', h}$ be a set of all ε'-significant states (see Definition E.1) at step h. Then we can rewrite this expectation as follows

$$\mathbb{E}_{\pi}[f(s_h, a_h)|s_1] = \sum_{a\in\mathcal{A}, s\in S_{\varepsilon', h}} d_h^{\pi}(s, a)f(s, a) + \sum_{a\in\mathcal{A}, s\notin S_{\varepsilon', h}} d_h^{\pi}(s, a)f(s, a).$$

For the first sum by Theorem E.2 we have $d_h^{\pi}(s, a) \leq 2SAH\mu_h(s, a)$, thus

$$\sum_{a\in\mathcal{A}, s\in S_{\varepsilon', h}} d_h^{\pi}(s, a)f(s, a) \leq 2SAH \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu_h(s, a)f(s, a) = 2SAH\mathbb{E}_{(s,a)\sim\mu_h}[f(s, a)].$$

For the second sum we apply $f(s, a) \leq B$ and the fact that for all states that are not ε'-significant under any policy $d_h^{\pi}(s) \leq \varepsilon'$:

$$\sum_{a\in\mathcal{A}, s\notin S_{\varepsilon', h}} d_h^{\pi}(s, a)f(s, a) \leq B \sum_{a\in\mathcal{A}, s\notin S_{\varepsilon', h}} d_h^{\pi}(s, a) = B \sum_{s\notin S_{\varepsilon', h}} d_h^{\pi}(s) \leq BS\varepsilon'.$$

$\qquad\square$

F. Faster Rates for Visitation Entropy

F.1. Algorithm description

Let us start from the description of the modified algorithm `RegEntGame`. It has a similar game-theoretical foundation as it aims at solving the following minimax game

$$\max_{d \in \mathcal{K}_p} \mathcal{H}_{\text{visit}}(d) = \max_{d \in \mathcal{K}_p} \min_{\bar{d} \in \mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)} = \min_{\bar{d} \in \mathcal{K}} \max_{d \in \mathcal{K}_p} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}.$$

As for usual `EntGame`, there are two players in the game. On the one hand, the min player, or forecaster player, tries to predict which state-action pairs the max player will visit to minimize $\text{KL}(d_h, \bar{d}_h)$. On the other hand, the max player, or sampler player, is rewarded for visiting state-action pairs that the forecaster player did not predict correctly.

We now describe the algorithm `RegEntGame` for MVEE. In this algorithm, we let a forecaster player and a sampler player compete for T episodes long. Let us first define the two players.

Forecaster-player The forecaster player remains exactly the same as for usual `EntGame` algorithm, see the corresponding section in the main text.

Regularized Sampler-player For the sampler player we exploit strong convexity of visitation entropy. The running time of the sampler player will be divided onto two stages, as it was done in `RL-Explore-Ent` algorithm.

Exploration phase Before the start of the game, the sampler-player uses some preprocessing time in order to explore the environment to learn a simple (non-markovian) preliminary exploration policy π^{mix}. This policy is used to construct an accurate enough estimates of transition probabilities. This policy is obtained, as in `RF-RL-Explore` and `RL-Explore-Ent`, by learning for each state-action pair (s, ah), a policy that reliably reaches this action pair (s, a) at step h. This can be done by running any regret minimization algorithm for the sparse reward function putting reward one at state s at step h and zero otherwise. The policy π^{mix} is defined as the mixture of the aforementioned policies.

Planning phase The second phase is starting during the running time of the algorithm. Since `RL-Explore-Ent` algorithm is essentially reward-free in a sense of working with an arbitrary reward functions.

For each episode t during the game we define the empirical regularized Bellman equations

$$\widehat{Q}_h^t(s,a) = \log \frac{1}{\bar{d}_h^{t+1}(s)} + \widehat{p}_h^t \widehat{V}_{h+1}^t(s,a)$$

$$\widehat{V}_h^t(s) = \max_{\pi \in \Delta_{\mathcal{A}}} \{\pi \widehat{Q}_h^t(s,a) + \mathcal{H}(\pi)\}, \tag{24}$$

where $\widehat{V}_{H+1}^t = 0$. The sampler player then follows $d^{\pi^{t+1}}$ where π^{t+1} is greedy with respect to the regularized Q-values, that is, $\pi_h^{t+1}(s) \in \arg\max_{\pi \in \Delta_A} \{\pi \widehat{Q}_h^t(s) + \mathcal{H}(\pi)\}$. This choice of sampler player will be clear in the analysis below.

Sampling rule At each episode t, the policy π^t of the sampler-player is used as a sampling rule to generate a new trajectory.

Decision rule After T episodes we output a non-Markovian policy $\widehat{\pi}$ defined as the mixture of the policies $\{\pi^t\}_{t \in [T]}$, that is, to obtain a trajectory from $\widehat{\pi}$ we first sample uniformly at random $t \in [T]$ and then follow the policy π^t. Note that the visitation distribution of $\widehat{\pi}$ is exactly the average $d^{\widehat{\pi}} = (1/T) \sum_{t \in [T]} d^{\pi^t}$.

Remark that the stopping rule of `RegEntGame` is deterministic and equals to $\tau = T$. The complete procedure is detailed in Algorithm 4.

Algorithm 4 `RegEntGame`
___
1: **Input:** Number of episodes T, number of exploration episodes N_0, number of transition samples N, prior counts n_0.
2: # Preliminary exploration
3: **for** $(s', h') \in \mathcal{S} \times [H]$ **do**
4:     Form rewards $r_h(s, a) = \mathbb{1}\{s = s', h = h'\}$.
5:     Run EULER (Zanette & Brunskill, 2019) with rewards r_h over N_0 iterates and collect all policies $\Pi_{s', h'}$.
6:     Modify $\pi \in \Pi_{s', h'} : \pi_{h'}(a|s') = 1/A$ for all $a \in \mathcal{A}$.
7: **end for**
8: Construct a uniform mixture policy π^{mix} over all $\{\Pi_{s, h} : (s, h) \in \mathcal{S} \times [H]\}$.
9: Sample N independent trajectories $\{z_n\}_{n \in [N]}$ using π^{mix} in the original MDP.
10: Construct from $\{z_n\}_{n \in [N]}$ the estimates $\widehat{p}_h$ as in (19).
11: **for** $t \in [T]$ **do**
12:     # Forecaster-player
13:     Update pseudo counts $\bar{n}_h^{t-1}(s, a)$ and predict $\bar{d}_h^t(s, a)$.
14:     # Sampler-player
15:     Compute π^t by regularized planning (24) with rewards $\log\left(1/\bar{d}_h^t(s)\right)$ and entropy regularization.
16:     # Sampling
17:     **for** $h \in [H]$ **do**
18:         Play $a_h^t \sim \pi_h^t(s_h^t)$
19:         Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$
20:     **end for**
21:     Update counts and transition estimates.
22: **end for**
23: Output $\widehat{\pi}$ the uniform mixture of $\{\pi^t\}_{t \in [T]}$.
___

F.2. Analysis

We first define the regrets of each players obtained by playing T times the games. For the forecaster-player, for any $\bar{d} \in \mathcal{K}$, we define

$$\mathfrak{R}_{\mathrm{Fore}}^T(\bar{d}) \triangleq \sum_{t=1}^T \sum_{h,s,a} \widetilde{d}_h^t(s, a) \left(\log \frac{1}{\bar{d}_h^t(s, a)} - \log \frac{1}{\bar{d}_h(s, a)} \right)$$

where $\widetilde{d}_h^t(s, a) \triangleq \mathbb{1}\{(s_h^t, a_h^t) = (s, a)\}$ is a sample from $d_h^{\pi^t}(s, a)$. Similarly for the sampler-player, for any $d \in \mathcal{K}_p$, we define a *regularized regret*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d) \triangleq \sum_{t=1}^T \left(\sum_{h,s,a} [d_h(s, a) - d_h^{\pi^t}(s, a)] \log \frac{1}{\bar{d}_h^t(s, a)} - \sum_{h,s} [d_h(s) \mathrm{KL}(\pi(s), \bar{\pi}_h^t(s)) - d_h^{\pi^t}(s) \mathrm{KL}(\pi_h^t(s), \bar{\pi}_h^t(s))] \right),$$

where corresponding policies are defined as $\pi_h(a|s) = d_h(s, a)/d_h(s)$ and $\bar{\pi}_h^t(a|s) = \bar{d}_h^t(s, a)/\bar{d}_h^t(s)$ for $d_h(s) = \sum_a d_h(s, a)$ and $\bar{d}_h^t(s) = \sum_a \bar{d}_h^t(s, a)$.

Recall that the visitation distribution of the policy π returned by `RegEntGame` is the average of the visitation distributions of the sampler-player $d_h^{\widehat{\pi}}(s, a) = \hat{d}_h^T(s, a) \triangleq (1/T) \sum_{t=1}^T d_h^{\pi^t}(s, a)$. We also denote by $\mathring{d}_h^T(s, a) \triangleq (1/T) \sum_{t=1}^T \widetilde{d}^t(s, a)$ the average of the 'sample' visitation distributions.

We now relate the difference between the optimal visitation entropy and the visitation entropy of the outputted policy $\widehat{\pi}$ to

the regrets of the two players. Indeed, using $\mathcal{H}(p) = \sum_{i \in [n]} p_i \log(1/q_i) - \mathrm{KL}(p, q)$ for all $(p, q) \in (\Delta_n)^2$ and

$$
\begin{aligned}
\mathrm{KL}(d_h^{\pi^\star}, \bar{d}_h^t) &= \sum_{s,a} d_h^{\pi^\star}(s, a) \log\left(\frac{d_h^{\pi^\star}(s, a)}{\bar{d}_h^t(s, a)} \right) \\
&= \sum_s d_h^{\pi^\star}(s) \log\left(\frac{d_h^{\pi^\star}(s)}{\bar{d}_h^t(s)} \right) + \sum_s d_h^{\pi^\star}(s) \sum_a \pi_h^\star(a|s) \log\left(\frac{\pi_h^\star(a|s)}{\bar{\pi}_h^t(a|s)} \right) \\
&\geq \sum_s d_h^{\pi^\star}(s) \,\mathrm{KL}(\pi_h^\star(s), \bar{\pi}_h^t(s)) .
\end{aligned}
$$

This inequality could be treated as a strong convexity of visitation entropy with respect to trajectory entropy since $\mathrm{KL}(d_h^{\pi^\star}, \bar{d}_h^t)$ is a *Bregman divergence* with respect to $\mathcal{H}_{\mathrm{visit}}$, and the final average of $\mathrm{KL}(\pi_h^\star(s), \bar{\pi}_h^t(s))$ is a Bregman divergence with respect to $\mathcal{H}_{\mathrm{traj}}$ (up to linearities).

Applying this inequality, we have

$$
\begin{aligned}
T\big(\mathcal{H}_{\mathrm{visit}}(d^{\pi^\star}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})\big) \leq{}& \sum_{t=1}^T \left(\sum_{h,a,s} d_h^{\pi^\star}(s, a) \log \frac{1}{\bar{d}_h^t(s, a)} - \sum_{h,s} d_h^{\pi^\star}(s) \,\mathrm{KL}(\pi_h^\star(s), \bar{\pi}_h^t(s)) \right) \\
&- \sum_{t=1}^T \widetilde{d}_h^t(s, a) \log \frac{1}{\mathring{d}_h^T(s, a)} + T\big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\hat{d}^T)\big) \\
\leq{}& \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star}) + \underbrace{\sum_{t=1}^T \sum_{h,s,a} \big(d_h^{\pi^t}(s, a) - \widetilde{d}_h^t(s, a)\big) \log \frac{1}{\bar{d}_h^t(s, a)}}_{\mathrm{Bias}_1} \\
&+ \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) + \underbrace{T\big(\mathcal{H}_{\mathrm{visit}}(\mathring{d}^T) - \mathcal{H}_{\mathrm{visit}}(\hat{d}^T)\big)}_{\mathrm{Bias}_2} .
\end{aligned}
$$

It remains to upper bound each terms separately in order to obtain a bound on the gap. Notably, only the sampler player result changes in comparison to `EntGame`.

F.3. Regret of the Sampler-Player

We start from introducing new notation. Let $\mathcal{M}_t = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h \in [H]}, \{r_h^t\}_{h \in [H]}, s_1)$ be a sequence of entropy-regularized MDPs where reward defined as $r_h^t(s, a) = \log(1/\bar{d}_h^t(s))$. Define $Q_h^{\pi,t}(s, a)$ and $V_h^{\pi,t}(s, a)$ as a action-value and value functions of a policy π on a MDP $\mathcal{M}_t$. Notice that the value-function of initial state in this case could be written as follows (see Appendix C)

$$
\begin{aligned}
V_1^{\pi,t}(s_1) &= \sum_{h,s,a} d_h^\pi(s, a) \log\left(\frac{1}{\bar{d}_h^t(s, a)} \right) - \sum_{h,s} d_h^\pi(s) \,\mathrm{KL}(\pi_h(s), \bar{\pi}_h^t(s)) \\
&= \sum_{h,s,a} d_h^\pi(s, a) \log\left(\frac{1}{\bar{d}_h^t(s)} \right) + \sum_{h,s} d_h^\pi(s) \mathcal{H}(\pi_h(s)),
\end{aligned}
$$

also see Appendix D for more exposition. Therefore, the regret for the sampler-player could be rewritten in the terms of the regret for this sequence of entropy-regularized MDPs

$$
\mathfrak{R}_{\mathrm{Samp}}^T(d^\pi) = \sum_{t=1}^T V_1^{\pi,t}(s_1) - V_1^{\pi^t,t}(s_1).
$$

We notice that our approach does not gives a regret minimizer algorithm in a classical sense, however analysis shows us that we can control the sum of policy error with respect to *any* reward function.

Lemma F.1. *Let $N_0 = \Omega\left(\frac{H^7 S^3 A \cdot \log^2(T+SA) \cdot L}{\varepsilon}\right)$ and $N = \Omega\left(\frac{H^6 S^3 A \log^2(T+SA) L^3}{\varepsilon}\right)$. Then with probability at least $1 - \delta/2$, the regret of the sampler player is bounded as*

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^\star}) \leq \varepsilon/2 \cdot T$$

after

$$\widetilde{\mathcal{O}}\left(\frac{H^8 S^4 A}{\varepsilon}\right)$$

episodes of pure exploration.

Proof. From Corollary E.6 under the choice of parameters $\lambda = 1, \kappa = 0$ and reward function $r_h^t(s,a) = \log(1/\bar{d}_h^t(s))$ for each iteration, that is bounded by $\log(T + SA)$, we have that for any reward function the sub-optimality gap is bounded by $\varepsilon/2$. The total number of episodes of pure exploration is equal to $N_0 SH + N$. $\qquad\square$

F.4. Proof of Theorem 5.1

We state the version of this theorem with all prescribed dependencies factors.

Theorem F.2. *Fix some $\varepsilon > 0$ and $\delta \in (0, 1)$. Then for $n_0 = 1$,*

$$N_0 = \Omega\left(\frac{H^7 S^3 A \cdot L^3}{\varepsilon}\right), \quad N = \Omega\left(\frac{H^6 S^3 A L^5}{\varepsilon}\right), \quad T = \Omega\left(\frac{H^2 S A L^3}{\varepsilon^2} + \frac{H^2 S^2 A^2 L^2}{\varepsilon}\right)$$

with $L = \log(SAH/\delta\varepsilon)$, the algorithm `RegEntGame` *is (ε, δ)-PAC. Its total sample complexity is equal to $SH \cdot N_0 + N + T$, that is,*

$$\tau = \widetilde{\mathcal{O}}\left(\frac{H^2 SA}{\varepsilon^2} + \frac{H^8 S^4 A}{\varepsilon}\right).$$

Proof. We start from writing down the decomposition defined in the beginning of the appendix

$$T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})) \leq \mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^{\star,\mathrm{VE}}}) + \mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) + \mathrm{Bias}_1 + \mathrm{Bias}_2.$$

By Lemma F.1 with probability at least $1 - \delta/2$ it holds

$$\mathfrak{R}_{\mathrm{Samp}}^T(d^{\pi^{\star,\mathrm{VE}}}) \leq \varepsilon T/2.$$

By Lemma B.1

$$\mathfrak{R}_{\mathrm{Fore}}^T(\mathring{d}^T) \leq HSA \log\left(\mathrm{e}(T+1)\right).$$

By Lemma B.6 with probability at least $1 - \delta/2$

$$\mathrm{Bias}_1 + \mathrm{Bias}_2 \leq 3\log(SAT)\left(\sqrt{TH \log(4/\delta)} + H\sqrt{SAT \log(3T)}\right).$$

By union bound all these inequalities hold simultaneously with probability at least $1 - \delta$. Combining all these bounds we get

$$T(\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})) \leq 3\log(SAT)\left(\sqrt{TH \log(4/\delta)} + H\sqrt{SAT \log(3T)}\right)$$
$$+ HSA \log(\mathrm{e}(T+1)) + \varepsilon T/2.$$

Therefore, it is enough to choose T such that $\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}})$ is guaranteed to be less than ε. In this case `RegEntGame` become automatically (ε, δ)-PAC. It is equivalent to find a maximal T such that

$$\varepsilon T/2 \leq 3\log(SAT)\left(\sqrt{TH \log(4/\delta)} + H\sqrt{SAT \log(3T)}\right) + HSA \log(\mathrm{e}(T+1))).$$

and add 1 to it. We start from obtaining a loose bound to eliminate logarithmic factors in T.

First, we assume that $T \geq 1$, thus $T + 1 \leq 2T$. Additionally, let us use inequality $\log(x) \leq x^\beta / \beta$ for any $x > 0$ and $\beta > 0$. We obtain

$$\varepsilon T \leq 48(SAT)^{1/8}\left(\sqrt{T^{3/2}H\log(4/\delta)} + H\sqrt{4SAT^{3/2}}\right) + 16/7 \cdot HSA(2\mathrm{e}T)^{7/8}$$

that could be relaxed as follows

$$\varepsilon T^{1/8} \leq 48(SA)^{1/8}(H\log(4/\delta)^{1/2} + H(SA)^{5/8} + 11HSA$$

thus we can define $\gamma = 8\log\big((48(SA)^{1/8}(H\log(4/\delta)^{1/2} + H(SA)^{5/8} + 11HSA)/\varepsilon\big) = \mathcal{O}(L)$ for which $\log(T) \leq \gamma$. Therefore

$$\varepsilon T/2 \leq 3(\log(SA) + \gamma)\sqrt{T}\left(\sqrt{H\log(4/\delta)} + \sqrt{SAH^2(\log(3) + L)}\right) + HSA(1 + 2\gamma).$$

Solving this quadratic inequality, we obtain the minimal required T to guarantee $\mathcal{H}_{\mathrm{visit}}(d^{\pi^{\star,\mathrm{VE}}}) - \mathcal{H}_{\mathrm{visit}}(d^{\widehat{\pi}}) \leq \varepsilon$. In particular,

$$T = \Omega\left(\frac{H^2 SAL^3}{\varepsilon^2} + \frac{H^2 S^2 A^2 L^2}{\varepsilon}\right).$$

$\square$

G. Deviation Inequalities

G.1. Deviation Inequality for Categorical Distributions

Next, we state the deviation inequality for categorical distributions by Jonsson et al. (2020, Proposition 1). Let $(X_t)_{t \in \mathbb{N}^\star}$ be i.i.d. samples from a distribution supported on $\{1, \ldots, m\}$, of probabilities given by $p \in \Delta_{m-1}$, where Δ_{m-1} is the probability simplex of dimension $m - 1$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e., for all $k \in \{1, \ldots, m\}$,

$$\widehat{p}_{n,k} \triangleq \frac{1}{n} \sum_{\ell=1}^{n} \mathbb{1}\{X_\ell = k\}.$$

Note that an element $p \in \Delta_{m-1}$ can be seen as an element of $\mathbb{R}^{m-1}$ since $p_m = 1 - \sum_{k=1}^{m-1} p_k$. This will be clear from the context.

Theorem G.1. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0, 1]$,*

$$\mathbb{P}(\exists n \in \mathbb{N}^\star,\ n\,\mathrm{KL}(\widehat{p}_n, p) > \log(1/\delta) + (m-1)\log(e(1 + n/(m-1)))) \leq \delta.$$

G.2. Deviation Inequality for Shannon Entropy

We denote by $\mathcal{H}(p)$ the (Shannon) entropy of $p \in \Delta_{m-1}$,

$$\mathcal{H}(p) \triangleq \sum_{k=1}^{m} p_k \log\left(\frac{1}{p_k}\right).$$

We will follow the ideas of Paninski (2003).

Theorem G.2. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0, 1]$*

$$\mathbb{P}\left[|\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geq \sqrt{\frac{2\log^2(n) \cdot \log(2/\delta)}{n}} + \left(\frac{(m-1)\log(e(1 + n/(m-1))) + 1}{n} \wedge \log(m)\right)\right] \leq \delta.$$

Moreover,

$$\mathbb{P}\left[\exists n : |\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geq \sqrt{\frac{2\log^2(n) \cdot (\log(2/\delta) + \log(n(n+1)))}{n}} + \left(\frac{m\log(e(1+n))}{n} \wedge \log(m)\right)\right] \leq \delta.$$

Proof. We start from application of McDiarmid's inequality to entropy by Antos & Kontoyiannis (2001). For all $p \in \Delta_{m-1}$ with probability at least $1 - \delta$ we have

$$\mathbb{P}\left[|\mathcal{H}(\hat{p}_n) - \mathbb{E}[\mathcal{H}(\hat{p}_n)]| \geq \sqrt{\frac{2\log^2(n) \cdot \log(2/\delta)}{n}}\right] \leq \delta.$$

To relate $\mathbb{E}[\mathcal{H}(\widehat{p}_n)]$ and $\mathcal{H}(p)$ we use the following observation

$$\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p) = -\mathrm{KL}(\widehat{p}_n, p) + \sum_{k:p_k>0} (\widehat{p}_{n,k} - p_k) \log(1/p_k),$$

therefore by taking expectation we have

$$\mathbb{E}[\mathcal{H}(\widehat{p}_n)] - \mathcal{H}(p) = -\mathbb{E}[\mathrm{KL}(\widehat{p}_n, p)].$$

In the following our analysis differs from (Paninski, 2003) since we obtain a direct estimate on the KL-divergence using Theorem G.1,

$$\mathbb{E}[n\,\mathrm{KL}(\widehat{p}_n, p)] = \int_0^\infty \mathbb{P}[n\,\mathrm{KL}(\widehat{p}_n, p) > t]\mathrm{d}t \leq (m-1)\log(e(1 + n/(m-1))) + \int_0^\infty e^{-t}\mathrm{d}t.$$

At the same time we have a trivial bound that concludes the first statement

$$\mathbb{E}[\mathcal{H}(\widehat{p}^n)] - \mathcal{H}(p) \leq \log(m).$$

To show the second statement of Theorem G.2, we apply the first part with $\delta'(n) = \delta/(n(n+1))$,

$$\mathbb{P}\left[|\mathcal{H}(\widehat{p}_n) - \mathcal{H}(p)| \geq \sqrt{\frac{2\log^2(n) \cdot \log(2/\delta'(n))}{n}} + \left(\frac{(m-1)\log(e(1 + n/(m-1))) + 1}{n} \wedge \log(m)\right)\right] \leq \frac{\delta}{n(n+1)},$$

thus by union bound over $n \in \mathbb{N}$ we conclude the statement. $\qquad\square$

G.3. Deviation Inequality for Sequence of Bernoulli Random Variables

Below, we state the deviation inequality for Bernoulli distributions by Dann et al. (2017, Lemma F.4). Let $\mathcal{F}_t$ for $t \in \mathbb{N}$ be a filtration and $(X_t)_{t\in\mathbb{N}^\star}$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_t = 1|\mathcal{F}_{t-1}) = P_t$ with P_t being $\mathcal{F}_{t-1}$-measurable and X_t being $\mathcal{F}_t$-measurable.

Theorem G.3. *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists n : \sum_{t=1}^n X_t < \sum_{t=1}^n P_t/2 - \log\frac{1}{\delta}\right) \leq \delta.$$

G.4. Deviation Inequality for Bounded Distributions

Below, we state the self-normalized Bernstein-type inequality by Domingues et al. (2021b). Let $(Y_t)_{t\in\mathbb{N}^\star}$, $(w_t)_{t\in\mathbb{N}^\star}$ be two sequences of random variables adapted to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. We assume that the weights are in the unit interval $w_t \in [0,1]$ and predictable, i.e. $\mathcal{F}_{t-1}$ measurable. We also assume that the random variables Y_t are bounded $|Y_t| \leq b$ and centered $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$. Consider the following quantities

$$S_t \triangleq \sum_{s=1}^t w_s Y_s, \quad V_t \triangleq \sum_{s=1}^t w_s^2 \cdot \mathbb{E}[Y_s^2|\mathcal{F}_{s-1}], \quad \text{and} \quad W_t \triangleq \sum_{s=1}^t w_s$$

and let $h(x) \triangleq (x+1)\log(x+1) - x$ be the Cramér transform of a Poisson distribution of parameter 1.

Theorem G.4 (Bernstein-type concentration inequality). *For all $\delta > 0$,*

$$\mathbb{P}\left(\exists t \geq 1, (V_t/b^2 + 1)h\left(\frac{b|S_t|}{V_t + b^2}\right) \geq \log(1/\delta) + \log(4e(2t+1))\right) \leq \delta.$$

The previous inequality can be weakened to obtain a more explicit bound: if $b \geq 1$ with probability at least $1 - \delta$, for all $t \geq 1$,

$$|S_t| \leq \sqrt{2V_t\log(4e(2t+1)/\delta)} + 3b\log(4e(2t+1)/\delta).$$

G.5. Deviation Inequalities for Expectation over Sampling Measure

In this section we describe two inequalities that shows the deviations of quantities of empirical transition kernels constructed by an independent samples from some base distribution over states. Let $\{z_k\}_{k\in[N]}$ be a set of independent trajectories using a fixed policy π in the original MDP: $\forall i \in [H] : a_i \sim \pi(s_i), s_{i+1} \sim p(s_i, a_i)$, and let $\mu_h(s,a) \triangleq d_h^\pi(s,a)$ be its state-action visitation distribution.

Using this data, we construct an estimates of transition probabilities $\{\widehat{p}_h\}_{h\in[H]}$: for each state-action-step triple (s,a,h) we define $n_h(s,a)$ as a number of visits in these N trajectories:

$$n_h(s,a) = \sum_{k=1}^N \mathbb{1}\{(s_h^k, a_h^k) = (s,a)\} \quad n_h(s'|s,a) = \sum_{k=1}^N \mathbb{1}\{(s_h^k, a_h^k, s_{h+1}^k) = (s,a,s')\},$$

where $z_k = (s_1^k, a_1^k, \dots, s_H^k, a_H^k, s_{H+1}^k)$, and then the model constructed in a usual way as a maximum likelihood estimate

$$\widehat{p}_h(s'|s,a) = \begin{cases} \frac{n_h(s'|s,a)}{n_h(s,a)} & n_h(s,a) > 0 \\ \frac{1}{S} & n_h(s,a) = 0 \end{cases}.$$

Lemma G.5. *Suppose that $\widehat{p}_h$ is the empirical transitions formed using N independent trajectories $\{z_k\}_{k\in[N]}$ sampled independently using policy π in the original MDP. Then we probability at least $1 - \delta$ the following holds*

$$\forall h \in [H] : \mathbb{E}_{(s,a)\sim\mu_h}\left[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\right] \leq \frac{12S^2 A \log^2(SN) \cdot \log(4SAH/\delta)}{N},$$

where $\mu_h(s,a) = d_h^\pi(s,a)$.

Proof. Define $n_h(s,a)$ be a number of samples from a fixed state-action pair s, a. Then by Theorem G.2 we have with probability at least $1 - \delta/2$ for all triples $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leq \frac{2\log^2(n_h(s,a)) \cdot \log(4SAH/\delta)}{n_h(s,a)} + \frac{S\log(S)\log(n_h(s,a))}{n_h(s,a)}.$$

From other point of view, we have a trivial upper bound $\log^2(S)$. Thus, we obtain

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leq \frac{2\log^2(Sn_h(s,a)) \cdot \log(4SAH/\delta) + S\log(S)\log(n_h(s,a))}{n_h(s,a)} \wedge 1.$$

Next we define the following event

$$\mathcal{E}^{\mathrm{cnt}}(\delta) = \left\{ \forall (s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H] : n_h(s,a) \geq \frac{N}{2}\mu_h(s,a) - \log(2SAH/\delta) \right\}.$$

By Theorem G.3 it holds with probability at least $1 - \delta/2$, then we can apply Lemma H.5 and obtain the following bound with probability at least $1 - \delta$ by union bound

$$(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2 \leq 4\frac{2\log^2(SN) \cdot \log(4SAH/\delta) + S\log(S)\log(N)}{(N\mu_h(s,a)) \vee 1}.$$

Thus, taking expectation we have

$$\mathbb{E}_{(s,a)\sim\mu_h}\left[(\mathcal{H}(\widehat{p}_h(s,a)) - \mathcal{H}(p_h(s,a)))^2\right] \leq \sum_{s,a} \frac{\mu_h(s,a)}{(N\mu_h(s,a)) \vee 1} \cdot 4\left(2\log^2(SN)\log(4SAH/\delta) + S\log(S)\log(N)\right)$$

$$\leq \frac{12S^2 A \cdot \log^2(SN) \cdot \log(4SAH/\delta)}{N}.$$

$\square$

Lemma G.6. *[Lemma C.2 by Jin et al. 2020] Suppose that $\widehat{p}_h$ is the empirical transitions formed using N independent trajectories $\{z_k\}_{k\in[N]}$ sampled independently using policy π in the original MDP. Then there is an absolute constant $C > 0$ such that with probability at least $1 - \delta$ the following holds for all $h \in [H]$*

$$\max_{G:\, \mathcal{S}\to[0,H\mathrm{R}_{\max}]} \max_{\nu:\, \mathcal{S}\to\mathcal{A}} \mathbb{E}_{(s,a)\sim\mu_h}\left[\left([\widehat{p}_{h'} - p_{h'}]G(s,a)\right)^2 \mathbb{1}\{\nu(s) = a\}\right] \leq \frac{CH^2\mathrm{R}_{\max}^2 S}{N}\log\left(\frac{AH\mathrm{R}_{\max}N}{\delta}\right),$$

where $\mu_h(s,a) = d_h^\pi(s,a)$.

H. Technical Lemmas

H.1. Entropy Properties

Lemma H.1. *For any reward-free MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h \in [H]}, s_1)$ the following holds*

$$\max_{d \in \mathcal{K}_p} \mathcal{H}_{visit}(d) = \max_{d \in \mathcal{K}_p} \min_{\bar{d} \in \mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}$$

$$= \min_{\bar{d} \in \mathcal{K}} \max_{d \in \mathcal{K}_p} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}$$

Proof. The first equality is due to

$$\mathcal{H}_{visit}(d) = \sum_{h=1}^{H} \mathcal{H}(d_h) = \sum_{h=1}^{H} \mathcal{H}(d_h) + \mathrm{KL}(d_h, d_h) = \min_{\bar{d} \in \mathcal{K}} \sum_{h=1}^{H} (\mathcal{H}(d_h) + \mathrm{KL}(d_h, \bar{d}_h)) = \min_{\bar{d} \in \mathcal{K}} \sum_{(h,s,a)} d_h(s,a) \log \frac{1}{\bar{d}_h(s,a)}$$

and the second equality uses Sion's theorem (Sion, 1958) with the fact that $\mathcal{K}, \mathcal{K}_p$ are compact convex sets and $(d, \bar{d}) \mapsto -\sum_{h,s,a} d_h(s,a) \log \bar{d}_h(s,a)$ is concave-convex. $\square$

Lemma H.2. *The trajectory entropy $\mathcal{H}_{traj}(\pi)$ has the following representation*

$$\mathcal{H}_{traj}(\pi) = \sum_{m \in \mathcal{T}} q^\pi(m) \log \frac{1}{q^\pi(m)} = \mathbb{E}_\pi \left[\sum_{h=1}^{H} \mathcal{H}(p_h(s_h, a_h)) + \mathcal{H}(\pi_h(s_h)) | s_1 \right].$$

Proof. By a definition of $q^\pi(m)$:

$$q^\pi(m) = \pi_1(a_1|s_1) \prod_{h=2}^{H} p_{h-1}(s_h|s_{h-1}, a_{h-1}) \pi_h(a_h|s_h),$$

thus

$$\sum_{m \in \mathcal{T}} q^\pi(m) \log \frac{1}{q^\pi(m)} = - \sum_{m \in \mathcal{T}} q^\pi(m) \left(\log(\pi_1(a_1|s_1) + \sum_{h=2}^{H} \log(p_{h-1}(s_h|s_{h-1}, a_{h-1}) + \log(\pi_h(a_h|s_h)) \right).$$

By rearranging the terms we have

$$\mathcal{H}_{traj}(\pi) = \sum_{h=1}^{H} \sum_{m \in \mathcal{T}} q_h^\pi(m) \left(\log \frac{1}{p_h(s_{h+1}|s_h, a_h)} + \log \frac{1}{\pi_h(a_h|s_h)} \right),$$

where under convention p_H is a deterministic transition to s_1. Marginalizing $q_h^\pi(m)$ over s_h, a_h, s_{h+1} we get

$$\mathcal{H}_{traj}(\pi) = \sum_{h=1}^{H} \sum_{s,a,s'} d^\pi(s,a) p_h(s'|s,a) \left(\log \frac{1}{p_h(s'|s,a)} + \log \frac{1}{\pi_h(a|s)} \right)$$

$$= \sum_{s,a} d_h^\pi(s,a) \sum_{h=1}^{H} \mathcal{H}(p_h(s,a)) + \sum_{s} d_h^\pi(s) \mathcal{H}(\pi_h(s)).$$

$\square$

Lemma H.3. *For any reward-free MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \{p_h\}_{h \in [H]}, s_1)$ and any policy π it holds*

$$\mathcal{H}_{traj}(q^\pi) \leq \mathcal{H}_{visit}(d^\pi) \leq H \mathcal{H}_{traj}(q^\pi).$$

Proof. The first inequality is a result of the non-negativity of mutual information, specifically the difference $\mathcal{H}_{\text{visit}}(d^\pi) - \mathcal{H}_{\text{traj}}(q^\pi)$ is equal to the Kullback-Leibler divergence between q^π and the product distribution $\otimes_{h=1}^{H} d_h^\pi$,

$$\mathcal{H}_{\text{visit}}(d^\pi) = \sum_{s,a,h} d_h^\pi(s,a) \log \frac{1}{d_h^\pi(s,a)}$$

$$= \sum_{s,a,h} \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\} q^\pi(m) \log \frac{1}{d_h^\pi(s,a)}$$

$$= \sum_{h=1}^{H} \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} q^\pi(m) \log \frac{1}{d_h^\pi(s_h,a_h)}$$

$$= \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} q^\pi(m) \log \frac{1}{\prod_{h=1}^{H} d_h^\pi(s_h,a_h)}$$

$$= \text{KL}(q^\pi, \otimes_{h=1}^{H} d_h^\pi) + \mathcal{H}_{\text{traj}}(q^\pi)$$

$$\geq \mathcal{H}_{\text{traj}}(q^\pi).$$

The second inequality is simply a consequence of the monotonicity of logarithm

$$\mathcal{H}_{\text{traj}}(q^\pi) = \sum_{m\in\mathcal{T}} q^\pi(m) \log \frac{1}{q^\pi(m)}$$

$$= \sum_{s,a,h} \frac{1}{H} \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\} q^\pi(m) \log\left(\frac{1}{q^\pi(m)}\right)$$

$$\geq \sum_{s,a,h} \frac{1}{H} \sum_{m=(s_1,a_1,\ldots,s_H,a_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s_h,a_h)\} q^\pi(m)$$

$$\cdot \log\left(\frac{1}{\sum_{m'=(s'_1,a'_1,\ldots,s'_H,a'_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s'_h,a'_h)\} q^\pi(m')}\right)$$

$$= \frac{1}{H}\mathcal{H}_{\text{visit}}(d^\pi),$$

where we used that $\log(1/x) = -\log(x)$ is monotonically decreasing, and

$$\sum_{m'=(s'_1,a'_1,\ldots,s'_H,a'_H)\in\mathcal{T}} \mathbb{1}\{(s,a)=(s'_h,a'_h)\} q^\pi(m') \geq q^\pi(m)$$

for any $m = (s_1, a_1, \ldots, s_H, a_H)$ such that $(s_h, a_h) = (s, a)$. $\qquad\square$

Lemma H.4. *In a deterministic MDP, i.e. the transition probability distributions are deterministic, the maximum trajectory entropy policy is the uniform policy over actions*

$$\pi_h^{\star,TE}(s,a) = 1/A.$$

Proof. The proof is straightforward by using the Bellman equations for MTEE. By induction over the step h we prove that $\pi_h^{\star,\text{TE}}(s,a) = 1/A$ and $V_h^\star(s) = (H+1-h)\log(A)$. Assume the induction hypothesis at step $h+1$. Then we get $Q_h^\star(s,a) = 0 + p_h V_{h+1}^\star(s,a) = (H-h)\log(A)$ which yields $\pi_h^{\star,\text{TE}}(s,a) = 1/A$ and $V_h^\star(s) = \log\left(\sum_{a\in\mathcal{A}} A^{H-h}\right) = (H+1-h)\log(A)$. The base case $h = H$ is immediate. $\qquad\square$

H.2. Counts to Pseudo-counts

Here we state Lemma 8 and Lemma 9 by Ménard et al. (2021).

Lemma H.5. *On event $\mathcal{E}^{\text{cnt}}$, for any $\beta(\delta,\cdot)$ such that $x \mapsto \beta(\delta,x)/x$ is non-increasing for $x \geq 1$, $x \mapsto \beta(\delta,x)$ is non-decreasing $\forall h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\forall t \in \mathbb{N}^\star, \; \frac{\beta(\delta, n_h^t(s,a))}{n_h^t(s,a)} \wedge 1 \leq 4\frac{\beta(\delta, \bar{n}_h^t(s,a))}{\bar{n}_h^t(s,a) \vee 1}.$$

Lemma H.6. *For $T \in \mathbb{N}^\star$ and $(u_t)_{t \in \mathbb{N}^\star}$, for a sequence where $u_t \in [0,1]$ and $U_t \triangleq \sum_{l=1}^{t} u_\ell$, we get*

$$\sum_{t=0}^{T} \frac{u_{t+1}}{U_t \vee 1} \leq 4\log(U_{T+1} + 1).$$

H.3. On the Bernstein Inequality

We restate here a Bernstein-type inequality by Talebi & Maillard (2018).

Lemma H.7 (Corollary 11 by Talebi & Maillard, 2018)**.** *Let $p, q \in \Delta_S$, where Δ_S denotes the probability simplex of dimension S. For all functions $f : \mathcal{S} \mapsto [0, b]$ defined on $\mathcal{S}$,*

$$pf - qf \leq \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p, q)} + \frac{2}{3}b\,\mathrm{KL}(p, q)$$

$$qf - pf \leq \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p, q)}\,.$$

where use the expectation operator defined as $pf \triangleq \mathbb{E}_{s \sim p} f(s)$ and the variance operator defined as $\mathrm{Var}_p(f) \triangleq \mathbb{E}_{s \sim p}\big(f(s) - \mathbb{E}_{s' \sim p} f(s')\big)^2 = p(f - pf)^2$.

Lemma H.8. *Let $p, q \in \Delta_S$ and a function $f : \mathcal{S} \mapsto [0, b]$, then*

$$\mathrm{Var}_q(f) \leq 2\mathrm{Var}_p(f) + 4b^2\,\mathrm{KL}(p, q)\,,$$

$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_q(f) + 4b^2\,\mathrm{KL}(p, q).$$

Lemma H.9. *For $p, q \in \Delta_S$, for $f, g : \mathcal{S} \mapsto [0, b]$ two functions defined on $\mathcal{S}$, we have that*

$$\mathrm{Var}_p(f) \leq 2\mathrm{Var}_p(g) + 2bp|f - g| \quad \text{and}$$

$$\mathrm{Var}_q(f) \leq \mathrm{Var}_p(f) + 3b^2\|p - q\|_1,$$

where we denote the absolute operator by $|f|(s) = |f(s)|$ for all $s \in \mathcal{S}$.

I. Additional Experiments

In this section we provide details on experiments and additional experiments. The code for experiments could be found by the following link: https://github.com/d-tiapkin/max-entropy-exploration.

First, we describe our baselines in details.

- Random policy $\pi_h(a|s) = 1/A$ for any $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$;

- Optimal MVEE policy. First we compute the solution to the following convex program

$$
\max_{d \in \mathcal{K}_p} \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} - \left(\frac{1}{H} \sum_{h=1}^{H} d_h(s,a) \right) \log \left(\frac{1}{H} \sum_{h=1}^{H} d_h(s,a) \right),
$$

  where $\mathcal{K}_p$ is a polytope of admissible visitation distributions defined in Section 2. This problem exactly corresponds to the setting visitation entropy by Hazan et al. (2019). Then we compute the optimal MVEE policy by normalization of the visitation distribution $\pi_h(a|s) = \frac{d_h(s,a)}{\sum_{a \in \mathcal{A}} d_h(s,a)}$.

- Optimal MTEE policy that was computed by solving regularized Bellman equations with $\lambda = 1$ and the entropy of transition kernels as rewards.

The last two baselines requires the knowledge of the true transition kernel $p_h(s'|s,a)$. Additionally, we perform experiments with RF-UCRL algorithm by Kaufmann et al. (2021) but the final visitation numbers were very close to the visitation numbers of EntGame algorithm and we do not report them. The data collection procedure for EntGame and UCBVI-Ent is following.

- For EntGame algorithm we report visitation counts of states of the algorithm during the learning. It is an appropriate choice since this counts corresponds to the empirical density $\mathring{d}_h^T(s,a) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{(s_h^t, a_h^t) = (s,a)\}$ which converges to $\widehat{d}_h^{\widehat{\pi}}(s,a) = \frac{1}{T} \sum_{t=1}^{T} d_h^{\pi^t}(s,a)$.

- For UCBVI-Ent algorithm we have to separate stages. At the first stage the algorithm interacts with environment to learn the final policy during N interactions with the environment, and all the plots present the visitations counts only for the final policy during another N interactions when the policy is fixed.

Remark I.1. We do not compare EntGame with MaxEnt and Toc-UCRL2 because of their similarity. The only difference in the these algorithms is the way how density estimation is performed. More precisely, in the MaxEnt algorithm the average of state-action visitation distributions of all the policies played is estimated from scratch at the end of each epoch with additional trajectories. In EntGame we also estimate this average but with all trajectories collected until now and without need of additional fresh trajectories. Thus, EntGame can be seen as a version of MaxEnt that reuses past trajectories instead of collecting new ones. The latter feature is helpful in practice but will not change the rate. That is why we decided not to include MaxEnt in the experiments. Similar comments hold for Toc-UCRL2.

Double Chain. The experiment described in Section 6 was preformed on Double Chain environment described by Kaufmann et al. (2021). This MDP consists of states $\mathcal{S} = \{0, \dots, L-1\}$, where L is the length of the chain, the actions $\mathcal{A} = \{l, r\}$ which corresponds to the transition to the left (action l) or to the right (action r). Additionally, while taking the actions, there is 10% probability of moving to the opposite direction. The agent starts at the middle of the chain $s_1 = (L-1)/2$. For experiments we run each algorithm to collect $N = 100000$ samples during episodes of horizon $H = 20$ and then report the average and confidence intervals over 48 random seeds.

In Section 6 we conclude that the exploration bonuses have the important role in the state visitations of UCBVI-Ent algorithm and make it close to RF-UCRL by Kaufmann et al. (2021). As an ablation study we preform the same experiments for algorithms without bonuses. The results are presented in Figure 2.

In particular, we see that EntGame algorithm without bonuses converges to the optimal MVEE policy in terms of the visitation densities that is slightly more spread than EntGame algorithm with bonuses. Notice that algorithm has its own exploration mechanism since the rewards are equal to $\log((t + SA)/(n^t(s,a) + 1))$ that is tightly connected to exploration bonuses. UCBVI-Ent algorithm also has its own exploration method that induced by soft-max policies. In particular,
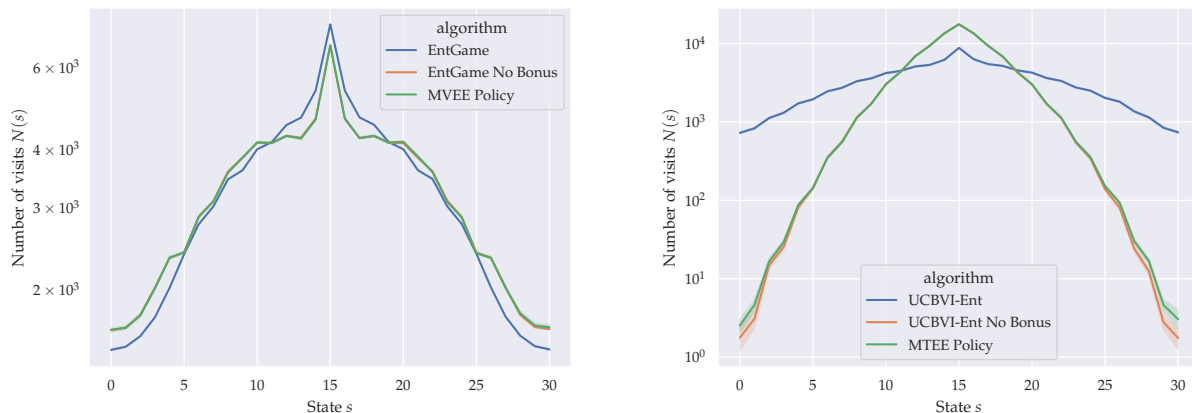
Figure 2. Number of state visits for $N = 100000$ samples in the Double Chain MDP for `EntGame` and `UCBVI-Ent` algorithms with and without bonuses.

soft-max policies are tightly connected to the Boltzmann exploration with step-size equal to 1 (Sutton, 1990). However, it is well-known (Cesa-Bianchi et al., 2017) that the Boltzmann exploration with fixed step-size did not provide efficient way to explore.
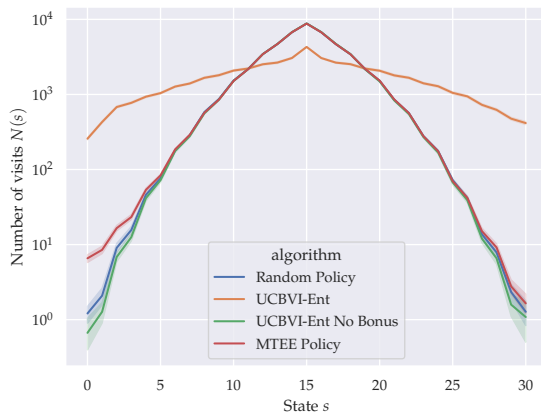

Figure 3. Number of state visits for $N = 50000$ samples in the Double Chain MDP with resampling for the `UCBVI-Ent` algorithm with and without bonuses.

Double Chain with Resampling To test the exploration mechanism of `UCBVI-Ent` without bonuses, we slightly modify Double Chain environment: now the visitation of the left end of the chain leads to uniform resampling over all states. The result is presented in Figure 3.

In particular, we observe that in this situation `UCBVI-Ent` algorithm without bonuses still acts like a random policy due to low exploration for the ends of the chain, whereas the optimal MTEE policy visits the left part of the chain more often. This observation imply that the additional exploration mechanism for `UCBVI-Ent` algorithm is required and the exploration problem in regularized MDPs is non-trivial.

GridWorld As an additional experiment to verify our findings we perform additional experiments on the environment Grid World as it presented in Figure 4. The state space is a set of discrete points in a 21×21 grid. For each state there are 4 possible actions: left, right, up or down, and for each action there is a 5% probability to move to the wrong direction. The initial state s_1 is the middle of the grid. For this experiment we use $N = 60000$ samples and report the average over 12
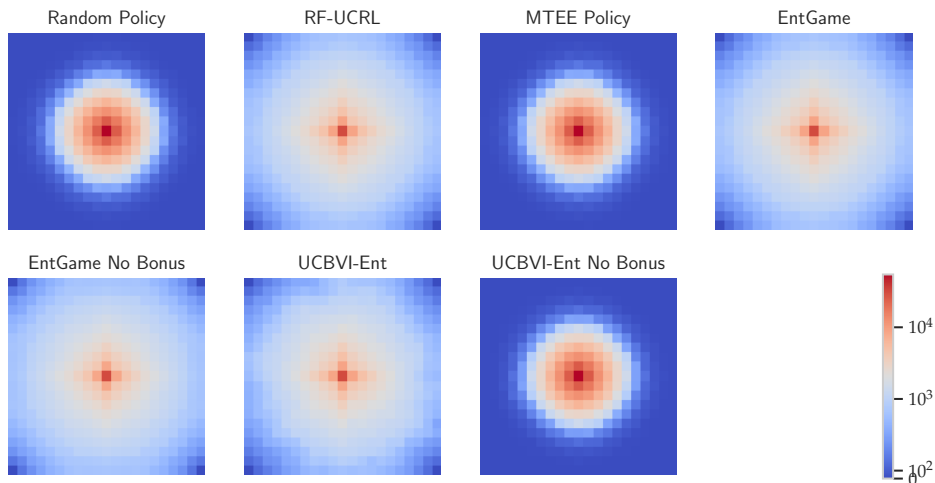
random seeds.

Figure 4. Number of state visits for $N = 60000$ samples in the GridWorld MDP for `EntGame` and `UCBVI-Ent` algorithms with and without bonuses.

Here we see the similar effect as on simpler environment: the `UCBVI-Ent` algorithm produces slightly less "spread" policy than `EntGame` or `RF-UCRL` algorithms. It is connected to the fact that in this case the limiting MTEE policy of `UCBVI-Ent` is again almost uniform policy due to near-deterministic structure of the MDP. Additionally, we remark that in this case the optimal MVEE policy is much harder to compute due to numerical issues, however, the `EntGame` algorithm without bonuses produces slightly more uniform distribution that coincides with the effect we observe during experiments with a Double Chain environment.