Tight Sample Complexity Bounds for Entropic Best Policy Identification

Amer Essakine AMER.ESSAKINE@ENS-PARIS-SACLAY.FR
ENS Paris Saclay

Claire Vernade CLAIRE.VERNADE@UTN.DE
University of Technology Nuremberg

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study best-policy identification for finite-horizon risk-sensitive reinforcement learning under the entropic risk measure. Recent work established a constant gap in the exponential horizon dependence between lower and upper bounds on the number of samples required to identify an approximately optimal policy. Precisely, known lower bounds scale in $\Omega(e^{|\beta|H})$ where H is the horizon of the MDP, while the state-of-the-art upper bound achieves at best $O(e^{2|\beta|H})$ (Mortensen and Talebi, 2025) using a generative model. We show that this extra exponential factor can be traced to overly loose concentration control for exponential utilities. To close this open gap, we revisit the analysis of this problem through a forward-model based algorithm building on KL-based exploration bonuses that we adapt to the entropic criterion. The improvement we get is due to two main novel technical innovations. We leverage the smoothness properties of the exponential utility to derive sharper concentration bounds, and we propose a new stopping rule that exploits further this tightness to obtain a sample complexity that matches the lower bound.

Keywords: Entropic risk measure, Risk-sensitive reinforcement learning, Best policy identification

1. Introduction

Risk-sensitive Reinforcement Learning (RL) studies the problem of learning a policy whose return distribution, rather than only its expectation, satisfies desirable properties, typically with respect to downside risk or tail events. In many applications ranging from finance to robotics (Charpentier et al., 2023; Polydoros and Nalpantidis, 2017), it may be more important for a decision maker to certify with high confidence that the return does not fall below a critical threshold than to maximize the expected reward.

A prominent example of a dynamically consistent risk criterion is the Entropic Risk Measure (Marthe et al., 2023; Howard and Matheson, 1972), which generalizes the expectation through an exponential utility function parameterized by a scalar $\beta \in \mathbb{R}$. The sign and magnitude of β control the risk sensitivity of the decision maker, interpolating between risk-seeking and risk-averse behaviors. Crucially, the entropic risk measure satisfies a risk-sensitive dynamic programming principle, yielding a Bellman-type recursion for finite-horizon Markov Decision Processes (MDPs) (Sutton and Barto, 2018). As shown by Marthe et al. (2023), this property in fact characterizes the entropic family among utility-based risk measures satisfying dynamic consistency.

In RL, the MDP is unknown and the learner must rely on sampled trajectories to estimate the model and plan under uncertainty. Mortensen and Talebi (2025) study this problem in a discounted infinite horizon setting and show that the sample complexity of identifying an ϵ-optimal policy must

depend exponentially on the horizon. More precisely, for a horizon H and entropic parameter β, they prove a lower bound showing that any algorithm requires at least

$$\Omega(\exp(|\beta|H))$$

samples (up to polynomial factors in the number of states S, actions A, and $1/\epsilon$) to identify an ϵ-optimal policy. In the same work, the authors derive an upper bound in $\exp(2|\beta|H)$ using access to a generative model and a careful application of the simulation lemma (see Section 2 for precise statements). While this gap may look benign at first, we show that this constant term in the exponential is due to a fundamentally loose handling of exponential utilities in exploration bonuses.

Beyond simply closing this open gap, we provide a new set of tools to design algorithms for risk-sensitive RL. Our algorithm builds on the forward-model framework (Menard et al., 2021) and tailors the exploration bonuses and stopping-time to the entropic risk measure. We also revisit previous analyses and extract a problem-dependent quantity, the *maximal achievable reward* $G_{\max}$, that gives a sharper dependence of the bounds on the MDP than directly using its upper bound H.

The main contributions of the paper are the following:

- First, we derive a lower bound for the best policy identification problem for the forward model. To the best of our knowledge, this is the first lower bound for BPI in this setting. The lower bound is expressed in terms of the maximal achievable reward $G_{\max}$ that we argue is the natural scale for the entropic risk measure.
- We present ENTROPIC-BPI, an algorithm that outputs a policy π that achieves (ε, δ)-PAC with optimal sample complexity up to logarithmic factors and up to $e^{2\max\{0,\beta\}}$ which is a constant for ε small enough. The exponential dependence is $e^{|\beta|G_{\max}(\mathcal{M})} \leqslant e^{|\beta|H}$ which removes the extra exponential factor with respect to previous work.

2. Risk-Sensitive Episodic Reinforcement Learning

We consider a finite-horizon MDP $\mathcal{M} = \left(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{p_h\}_{h=1}^H\right)$ where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces respectively, $H \in \mathbb{N}$ is the fixed horizon length, $p_h(\cdot \mid s, a)$ is the transition kernel at step h for each $h \in \{1, \dots, H\}$, and $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function at step h, which we assume to be bounded in $[0, 1]$.

In the learning problem, the transition kernels $\{p_h\}_{h=1}^H$ are unknown to the learner. Through interactions, the learner must devise a policy π that is a (possibly non-stationary) mapping prescribing which action to take in each state and step, with the goal of maximizing a suitable performance criterion. In the literature, this problem is addressed under two distinct interaction models.

- **Generative model** (Azar et al., 2012; Kearns et al., 2002): In the generative model (also called a sample oracle), the learner has access to an oracle such that for any query $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times \{1, \dots, H\}$, the oracle returns a sample (r, s') where $r = r_h(s, a)$ and $s' \sim p_h(\cdot \mid s, a)$. Each call to the oracle is independent of the past, and the learner may choose its queries adaptively based on all previously observed samples.
- **Online (forward) model** (Dann and Brunskill, 2015; Strehl et al., 2009): In the forward interaction model, the learner interacts with the MDP over episodes. At the beginning of each episode t, an initial state s_1^t is drawn from a fixed initial distribution μ_1 on $\mathcal{S}$. At each

step $h \in \{1, \dots, H\}$ of episode t, the learner observes the current state s_h^t, selects an action $a_h^t \in \mathcal{A}$ according to some policy π_t, then receives a reward $r_h(s_h^t, a_h^t)$ and the environment goes to the next state $s_{h+1}^t \sim p_h(\cdot \mid s_h^t, a_h^t)$.

In this work, we focus exclusively on the online (forward) interaction model. Moreover, rather than the standard risk-neutral objective where we optimize the total expected reward along the trajectory, we adopt a risk-sensitive performance criterion based on the entropic risk measure which reflects the agent's attitude toward uncertainty, allowing for both risk-seeking and risk-averse behavior.

Policy & value functions A deterministic policy π is a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ for all $h \in \{1, ..., H\}$, where every π_h maps each state to a single action. Let $(S_h, A_h)_{h=1}^H$ be the random trajectory induced by the policy $(\pi_h)_{h\in\{1,\dots,H\}}$ and the MDP dynamics, i.e., $A_h = \pi_h(S_h)$ and $S_{h+1} \sim p_h(\cdot \mid S_h, A_h)$ starting from a fixed initial state s_1[1]. Define the (random) cumulative return from step h by:

$$R_h^\pi = \sum_{i=h}^{H} r_i(S_i, A_i)$$

The entropic value function of π, denoted by V_h^π is defined as:

$$V_h^\pi(s) \triangleq \frac{1}{\beta} \log \mathbb{E}\left[\exp\left(\beta \sum_{i=h}^{H} r_i(s_i, a_i)\right) \mid S_h = s\right]$$

where $a_i \triangleq \pi_i(s_i)$ and $s_{i+1} \sim p_i(\cdot \mid s_i, a_i)$

The optimal entropic value functions are defined as $V_h^\star(s) \triangleq \sup_\pi V_h^\pi(s)$ for $h \in [H]$. As for the expected return, both V_h^π and $V_h^\star$ satisfy the Bellman equation (Borkar and Meyn, 2000; Sutton and Barto, 2018; Marthe et al., 2023) and can thus be computed efficiently.

To simplify notation, we introduce a couple of operators that allow us to write concisely the effect of transition kernels or policies on our functionals of interest. For any bounded function $f : \mathcal{S} \to \mathbb{R}$, we denote by $(p_h f)(s, a) \triangleq \mathbb{E}_{S'\sim p_h(\cdot|s,a)}[f(S')]$ the action of the Markov kernel p_h on f. For any function $g : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and (possibly stochastic) policy π_h, we write $(\pi_h g)(s) \triangleq \mathbb{E}_{A\sim\pi_h(\cdot|s)}[g(s, A)]$ for the composition with the policy at step h. Finally, we introduce a useful quantity that appears naturally in the analysis of the sample complexity.

Definition 1 (Maximum achievable reward in a trajectory) *For an episodic MDP $\mathcal{M}$, define the maximal achievable return (a deterministic quantity depending only on $\mathcal{M}$) as*

$$G_{\max}(\mathcal{M}) = \sup_\pi \sup_\omega R_1^\pi(\omega)$$

where ω covers all sources of randomness (initial state, transitions, policy's own randomization).

Intuitively, $G_{\max}(\mathcal{M})$ is the largest total reward that can occur in a single episode under the most favorable sequence of outcomes. We will express both the lower bound and the upper bound in the quantity $G_{\max}(\mathcal{M})$ rather than the horizon in previous works. It is the natural scale for the

1. As explained in (Fiechter, 1994), if the first state is sampled randomly as $s_1 \sim p_0$, we simply add an artificial first state s_0 such that for any action a, the transition probability is defined as the distribution $p_0(s_0, a) \triangleq p_0$. This augments the state space by one and the horizon by one, and the bounds carry over with only this constant-size modification.

entropic risk measure objective where the hardness comes from the exponential amplification of rewards. Note that we have the inequality $G_{\max}(\mathcal{M}) \leqslant H$ so our results can also be expressed (more loosely) as a function of the horizon.

Empirical MDP. Let $(s_h^t, a_h^t, s_{h+1}^t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ be a tuple observed by the algorithm at step h of episode t. For any $h \in \{1, ..., H\}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, define the visitation counts

$$n_h^t(s,a) \triangleq \sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i) = (s,a)\} \qquad n_h^t(s,a,s') \triangleq \sum_{i=1}^{t} \mathbb{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}$$

These quantities induce the empirical transition probabilities

$$\widehat{p}_h^t(s'|s,a) \triangleq \begin{cases} \dfrac{n_h^t(s,a,s')}{n_h^t(s,a)} & \text{if } n_h^t(s,a) > 0, \\ \dfrac{1}{|\mathcal{S}|} & \text{otherwise} \end{cases}$$

We denote $\overline{n}_h^t(s,a) = \mathbb{E}[n_h^t(s,a)]$ the expected number of visits, which is called the pseudo-counts.

Best Policy Identification (BPI) under entropic risk Our objective is to identify a (near) optimal policy with high probability. In each episode t, the agent follows a policy π^t (the *sample rule*) based only on the information collected up to and including episode $t-1$. At the end of each episode, the agent can decide to stop collecting data according to a *stopping rule* (we denote by τ its random stopping time) and outputs a guess $\hat{\pi}$ for the optimal policy.

A BPI algorithm is therefore made of a triple $\left((\pi^t)_{t\in\mathbb{N}}, \tau, \hat{\pi}\right)$. The goal is to build an (ε, δ)-PAC algorithm according to the following definition, for which the sample complexity, that is the number of exploration episodes τ, is as small as possible.

Definition 2 (PAC algorithm for BPI) *An algorithm is (ε, δ)-PAC for best policy identification if it returns a policy $\hat{\pi}$ after some number of episodes τ that satisfies*

$$\mathbb{P}\left(V_1^\star(s_1) - V_1^{\hat{\pi}}(s_1) \leqslant \varepsilon\right) \geqslant 1 - \delta$$

Prior bounds. For the entropic risk measure, the best policy identification problem is harder[2] because it magnifies tail outcomes—inflating high rewards when $\beta > 0$ and penalizing low rewards when $\beta < 0$. Moreover, as this line of work is relatively recent and most of the results aim to establish regret bounds, the results on PAC bounds remain scarce; to our knowledge, there are no reward-free algorithms (Jin et al., 2020; Kaufmann et al., 2021) with theoretical guarantees specifically tailored to this criterion.

Sample complexity with the generative model. In the discounted infinite-horizon, Mortensen and Talebi (2025) proved a lower bound on the number of oracle calls needed for an $\varepsilon-$optimal policy. In particular, they showed that an exponential dependency on the risk parameter β and on the horizon H is unavoidable: Any algorithm that outputs an ε-optimal policy with probability at least

2. In the sense that it admits higher lower bounds, see Sec. 3

$1-\delta$ must make at least $\Omega\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}-3}{|\beta|^2}\log\left(\frac{S}{c_2\delta}\right)\right)$ calls to the oracle. They also provided two model-based algorithms with respective sample complexity $\tilde{\mathcal{O}}\left(\frac{SA\gamma^2}{c_1(1-\gamma)^4\varepsilon^2}\frac{\left(e^{2|\beta|\frac{1}{1-\gamma}}-1\right)^2}{|\beta|^2}\log\left(\frac{SA}{c_2\delta}\right)\right)$ This gives the first explicit sample complexity characterization for the entropic risk measure objective with a generative model. It can be mapped back to our finite-horizon setting using $H=\frac{1}{1-\gamma}$, the effective horizon.

Forward model: There are no known BPI sample-complexity bounds for the entropic risk measure in this model. However, there exist bounds on the regret, a more forgiving metric that sums $V^{\pi_t}-V^*$ over episodes that can be connected back to BPI as we explain next. The first non-asymptotic results are for model-free optimistic algorithms RSVI/RSQ (Fei et al., 2020), establishing $\tilde{\mathcal{O}}\big(\lambda(|\beta|H^2)\sqrt{H^3S^2AT}\big)$ (RSVI) and $\tilde{\mathcal{O}}\big(\lambda(|\beta|H^2)\sqrt{H^4SAT}\big)$ (RSQ) regret with $\lambda(u)=(e^{3u}-1)/u$. Fei et al. (2021) introduce the exponential Bellman equation and a doubly-decaying bonus, removing an extra $e^{|\beta|H^2}$ factor and yielding a regret of $\tilde{\mathcal{O}}\Big(\frac{e^{|\beta|H}-1}{|\beta|H}\sqrt{H^4S^2AK}\Big)$. More recently, Hu and Leung (2023) adapt UCB-ADVANTAGE (Zhang et al., 2020) to the exponential-utility setting and derive a worst-case problem-independent regret bound $\tilde{\mathcal{O}}\Big(\frac{e^{|\beta|H}-1}{|\beta|}\sqrt{H^2SAK}\Big)$ along with a tighter problem-dependent bound that matches the information-theoretic lower bound up to logarithmic factors on a class of MDPs.

These results are not directly related to our problem, though in general there is a connection between controlling regret and BPI sample complexity. As noticed by Jin et al. (2020), a straightforward approach to convert the regret bound to a sample complexity bound is to output a random policy among the sequence of policies returned by the regret algorithm. This idea was later shown to be outperformed (for the expected return) by Kaufmann et al. (2021), and the same applies in the case of entropic risk measures. It can be easily checked that a naive conversion of the best regret upper bounds above would yield a sample complexity in $e^{2|\beta|H}$ and a dependence in $\frac{1}{\delta^2}$. This upper bound can thus be considered the best achievable sample complexity so far for this problem.

3. Lower bound

The case $\beta\neq 0$ is fundamentally harder than the risk-neutral case $\beta=0$. Under the entropic criterion, trajectories are reweighted exponentially according to their cumulative return: for $\beta>0$, rare high-return trajectories can dominate the objective, while for $\beta<0$, rare low-return trajectories can dominate it. Thus, unlike in the expected-return setting, rare events do not contribute only proportionally to their probability of occurrence; their contribution is amplified by a factor exponential in the return. Therefore, if achieving near-optimal performance requires hitting a "hard" state–action–time triple $(s^\star,a^\star,h^\star)$ that is reached with tiny probability, the learner must repeatedly experience these rare transitions to accurately evaluate their contribution since a small number of tail episodes can dominate the entropic objective. As a result, learning becomes exponentially hard in the horizon H and the risk parameter β.

Theorem 3 *Fix $S\geqslant 6, A\geqslant 2, H\in\mathbb{N}, \beta\in\mathbb{R}^\star$, and $\delta\in(0,1/16)$ and ε small enough (See condition (21)). There exists an MDP $\mathcal{M}_0$ with S states, A actions, horizon H, and rewards in $[0,1]$ and nonstationary transitions such that for every algorithm $\mathcal{A}$ that outputs a policy $\hat{\pi}$ that is*

(ε, δ)-PAC for the entropic risk measure after sampling τ trajectories we have:

$$\mathbb{E}_{\mathcal{M}_0}[\tau] \geqslant \frac{1}{1650} \frac{(e^{|\beta| G_{\max}(\mathcal{M}_0)} - 1)^2}{e^{|\beta| G_{\max}(\mathcal{M}_0)}} \frac{e^{2\min\{\beta,0\}\varepsilon} SAH}{(e^{|\beta|\varepsilon} - 1)^2} \log\left(\frac{1}{\delta}\right)$$

Proof [Sketch of proof] We follow the hard episodic, stage-dependent MDP class introduced by Domingues et al. (2021). At a high level, these instances behave like a single multi-armed bandit with $\Theta(HSA)$ arms, where an "arm" corresponds to a triple (time, leaf, action). The agent starts in a waiting state s_w and can play a special action a_w to remain in s_w up to some stage $\bar{H}$; once it stops waiting (or after $\bar{H}$), it is forced to leave s_w and then traverses a full A-ary tree deterministically (each action selects the corresponding child), reaching a leaf after d steps. From a leaf state $s_i \in \mathcal{L}$, at stage h and action a, the process transitions to an absorbing good state s_g with probability $p_h(s_g|s_i, a)$ and to an absorbing bad state s_b otherwise. Rewards are obtained only in s_g for the rest of the horizon.

Consequently, achieving the optimal value requires (i) leaving s_w so as to hit the correct stage $h^\star$ at the leaves, (ii) reaching the correct leaf $s_{\ell^\star}$ (via the deterministic actions along the tree), and (iii) playing the correct action $a^\star$ at that leaf; only then does the probability of reaching s_g increase from p_- to p_+. We denote $u = (h^\star, l^\star, a^\star)$ for the rest of the proof.

The lower bound proof then compares a reference instance $\mathcal{M}_0$ (where no unique action is optimal) to instances $\mathcal{M}_u$ (where only one triplet has the favorable probability p_+), and applies standard change-of-measure arguments to show that distinguishing these close Bernoulli transition models forces a large number of episodes visiting the special triplet.

The construction of p_- for the entropic risk measure differs from the risk-neutral setting (where $p_- = \frac{1}{2}$) to reflect the hardness induced by the entropic risk measure:

- For $\beta > 0$, the entropic criterion overweights rare high-return trajectories, so we make success (reaching the good state) rare by choosing $p_- \sim e^{-\beta H}$
- For $\beta < 0$, the entropic criterion is especially sensitive to adverse tail events, so we instead make failure rare by choosing $p_- \sim 1 - e^{-|\beta| H}$

Then, we choose the gap $\Delta = p_+ - p_-$ so that, in instance M_u, any policy that does not identify the special triplet is ε-suboptimal for the entropic objective. A change of measure argument (Kaufmann et al., 2016) then gives the lower bound. The full proof together with details on the MDP construction can be found in Appendix C ■

To the best of our knowledge, this establishes the first lower bound for Best Policy Identification (BPI) in the forward model setting with non-stationary transitions.

When $|\beta|$ goes to 0, the lower bound approaches:

$$\Omega\left(\frac{G_{\max}(\mathcal{M}_0)^2}{\varepsilon^2} SAH \log\left(\frac{1}{\delta}\right)\right) = \Omega\left(\frac{SAH^3}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right)$$

The second equality holds because, in the construction of $\mathcal{M}_0$, the maximum return $G_{\max}(\mathcal{M}_0)$ scales linearly with H. Thus, we recover the standard lower bound for the risk-neutral case.

Moreover, for sufficiently small ε, the bound simplifies to:

$$\Omega\left(\frac{SAH}{\varepsilon^2} e^{|\beta| G_{\max}(\mathcal{M}_0)}\right)$$

This matches the lower bound derived by (Mortensen and Talebi, 2025) in the generative model setting up to an additional factor of H. This extra factor is inherent to the non-stationary transition dynamics of the finite-horizon setting. Another difference is that the exponential dependence in our bound scales with $G_{\max}(\mathcal{M}_0)$ rather than the horizon H (or the effective horizon $\frac{1}{1-\gamma}$). This distinction is intuitive: since the hardness of the entropic risk measure comes from exponentially reweighting trajectory returns, the exponential dependence is naturally governed by the maximum cumulative reward ($G_{\max}$) rather than the length of the episode.

4. Algorithm and Matching Upper Bound

We now present an algorithm whose sample complexity matches the lower bound (Theorem 3) up to logarithmic factors. A central difficulty with the entropic risk measure is that, unlike the risk-neutral objective, it is neither additive nor sub-additive, which complicates both algorithm design and analysis. In particular, standard UCB-style approaches rely on concentration inequalities for additive returns, whereas the entropic criterion involves a log-moment generating function and does not directly fit into standard risk-neutral concentration frameworks.

A common workaround is to use the Lipschitz continuity of the logarithm to reduce the analysis to risk-neutral quantities; however, this can yield coarse bounds and, more importantly, may break the dynamic-programming structure. As observed by Fei et al. (2021), losing a Bellman-type recursion can cause error terms to compound over the horizon, leading to substantially worse dependence on H in the order of $e^{2\beta H^2}$. Similarly to Fei et al. (2021), we instead work directly in the exponential space induced by the criterion. This restores a Bellman recursion for the exponentiated value functions and enables sharper control of uncertainty. In particular, Lemma 25 derives Bellman-type variance identities in this space, which are unavailable in the original entropic-value space due to the lack of sub-linearity[3]. This identity highlights that the exponential parameterization is the natural domain for controlling uncertainty under the entropic criterion.

For a policy π, the exponential transforms of the value and the state-value function are:

$$Z_h^\pi(s) \triangleq \exp(\beta V_h^\pi(s)), \qquad U_h^\pi(s,a) \triangleq \exp(\beta Q_h^\pi(s,a)).$$

We introduce two novel techniques. We adapt the KL-based exploration bonuses introduced in (Menard et al., 2021) to the entropic criterion to obtain bonuses with variance-sensitive control in the exponential space defined by these exponential transforms. We also propose an entropic stopping rule that yields sharper horizon dependence, improving over bounds that incur an additional factor of order $H^2 e^{|\beta| H}$ (Mortensen and Talebi, 2025).

Similarly to Azar et al. (2017), Zanette and Brunskill (2019) and Menard et al. (2021) we define optimistic (and pessimistic, see (13)) state-value functions for $\beta > 0$ on the exponential transform $U_h^\star$ of $Q_h^\star$ (see (19) for $\beta < 0$): fix $\widetilde{U}_{H+1}^t(s,a) = 1$ and recursively,

$$\widetilde{U}_h^t(s,a) = \min\left\{ e^{\beta(H-h-1)}, e^{\beta r_h(s,a)} \left[\widehat{p}_h^t \widetilde{Z}_h^t(s,a) + b_h^t(s,a) + \frac{1}{H} \widehat{p}_h^t \left(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t \right)(s,a) \right] \right\}, \tag{1}$$

3. As shown by Rowland et al. (2019), the variance is a Bellman-closed statistic and can be computed by dynamic programming. We derive this result for the variance of exponential utilities

where $\widetilde{Z}_h^t$ and $\underset{\sim}{Z}_{h+1}^t(s,a)$ denote respectively the optimistic and pessimistic exponential value functions[4]: (see again (13) for pessimistic versions)

$$\widetilde{Z}_{H+1}^t(s) = 1, \quad \text{and recursively} \quad \widetilde{Z}_h^t(s) = \max_{a\in\mathcal{A}} \widetilde{U}(s,a), \tag{2}$$

and the bonus term is defined as:

$$b_h^t(s,a) = 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \tag{3}$$

with the exploration rates α, α^* (see Eq. (11) for exact definitions) being of the form $n, \delta \mapsto O(\log(SAH/\delta) + \log(n))$ and coming directly from the Bernstein inequality. As stated before, our algorithm acts greedily on this optimistic value function. Specifically, at step h in episode t, given state $s_{h,t}$, ENTROPIC-BPI executes

$$a_{h,t} = \pi_t(s_{h,t}) = \arg\max_a \widetilde{U}_h^t(s_{h,t}, a) \tag{4}$$

Entropic certificate Following related work, we call certificate an upper bound on the width of the confidence interval for a policy π. We define it by backward recursion:

$$\pi_{h+1}^{t+1}G_h^t(s) = \min\left\{e^{\beta(H-h-1)}, e^{\beta r_h(s,\pi_h^{t+1}(s))}\left(3b_h^t(s,\pi_h^{t+1}(s)) + \left(1+\frac{3}{H}\right)\widehat{p}_h^t(\pi_{h+1}^{t+1}G_{h+1}^t)(s,\pi_h^{t+1}(s))\right)\right\} \tag{5}$$

with terminal $G_{H+1}^t \equiv 0$. In particular, on the high-probability good event, the performance gap of the greedy policy π_{t+1} is controlled by the start-state certificate. Concretely, lemma 12 shows that with high probability $1-\delta$:

$$Z_1^\star(s_1) - Z_1^{\pi_{t+1}}(s_1) \leqslant \widetilde{Z}_1(s_1) - Z_1^{\pi_{t+1}}(s_1) \leqslant \pi_{t+1,1}G_1^t(s_1) \tag{6}$$

Stopping rule We derive a stopping criterion based on the certificate G_h^t by relating the value function gap to the ratio of partition functions. The difference in value functions can be expressed in the exponential space as:

$$\begin{aligned}(V_1^\star - V_1^{\pi^{t+1}})(s_1) &= \frac{1}{\beta}\log\left(\frac{Z_1^\star}{Z_1^{\pi^{t+1}}}\right)(s_1) \\ &= \frac{1}{\beta}\log\left(1 + \frac{Z_1^\star - Z_1^{\pi^{t+1}}}{Z_1^{\pi^{t+1}}}\right)(s_1).\end{aligned}$$

To ensure the policy is ε-optimal, it suffices to satisfy at stopping time τ:

$$\pi_1^t G_1(s_1) \leqslant \left(e^{\beta\varepsilon} - 1\right) Z_1^{\pi^t}(s_1) \tag{7}$$

However, $Z_1^{\pi^t}$ is unknown as it relies on the true dynamics. We therefore substitute it with a computable lower bound. Using (6), we have:

$$Z_1^{\pi_{t+1}}(s_1) \geqslant \widetilde{Z}_1^t(s_1) - \pi_{t+1,1}G_1^t(s_1).$$

4. We use upper and lower tilde accents to indicate optimism and pessimism.

Substituting this lower bound into the optimality condition (7) yields a stronger, computable stopping rule:

$$\pi_1^t G_1(s_1) \leqslant \left(e^{\beta\varepsilon} - 1\right)\left(\widetilde{Z}_1^t(s_1) - \pi_1^t G_1(s_1)\right).$$

Solving for $\pi_1^t G_1(s_1)$, we obtain the final stopping criterion:

$$\pi_1^t G_1(s_1) \leqslant \frac{e^{\beta\varepsilon} - 1}{e^{\beta\varepsilon}} \widetilde{Z}_1^t(s_1) \tag{8}$$

where both $\pi_1^t G_1(s_1)$ and $\widetilde{Z}_1^t(s_1)$ can be computed efficiently by dynamic programming.

Insight on the stopping rule To guarantee ε-optimality, we need $\pi_1 G_1(s_1)/Z_1^{\pi_1^t}(s_1)$ to be smaller than a threshold $e^{\beta\varepsilon} - 1$. Our analysis reveals that (see proof below)

$$\pi_1 G_1(s_1)/Z_1^{\pi_1^t}(s_1) \lesssim \mathcal{O}\left(\sqrt{\frac{\mathrm{Var}\left(e^{\beta R_1^\pi}\middle|S_1 = s_1\right)}{\mathbb{E}\left[e^{\beta R_1^\pi}\middle|S_1 = s_1\right]^2}\mathbb{E}^\pi\left[\sum_{h=1}^{H}\frac{1}{n_h^t(s,a)}\middle|s_1\right]}\right) \tag{9}$$

The bound consists of a decreasing visitation term and the constant $\frac{\mathrm{Var}\left(e^{\beta R_1^\pi}\middle|S_1=s_1\right)}{\mathbb{E}\left[e^{\beta R_1^\pi}\middle|S_1=s_1\right]^2}$, which governs the asymptotic rate in contrast to $\mathrm{Var}(\mathrm{R}_1^\pi|\mathrm{S}_1 = \mathrm{s}_1)$ in the risk-neutral setting.

It is insightful to make a short comment on the connection of this quantity with the probability space we are working with. Let us denote $\mathbb{P}^\pi$ the probability distribution of a random trajectory $(s_1, a_1, ..., s_H, a_H)$ in the MDP, and consider the tilted law $\mathbb{P}_\beta^\pi$ defined by the Radon-Nykodim derivative $\frac{d\mathbb{P}_\beta^\pi}{\mathbb{P}^\pi}(\omega) = \frac{e^{\beta R_1^\pi(\omega)}}{Z_1^\pi(s_1)}$. We can see the tilted measure as the law of a trajectory on a twisted version of the original MDP that biases transitions towards states with high future exponential return. It can be easily computed that:

$$\frac{\mathrm{Var}\left(e^{\beta R_1^\pi}\middle|S_1 = s_1\right)}{\mathbb{E}\left[e^{\beta R_1^\pi}\middle|S_1 = s_1\right]^2} = \chi^2(\mathbb{P}_\beta^\pi, \mathbb{P}^\pi)$$

In other words, the constant leading the convergence speed in the upper bound (9) is the χ_2 divergence mismatch between the tilted trajectory distribution and the nominal trajectory distribution: It measures the extent to which optimizing the entropic objective amounts to learning under an implicit twisted dynamics that over-samples trajectories with high exponentiated reward. This mismatch is precisely what gets introduced by maximizing the entropic risk measure and what drives the extra constant $\frac{e^{\beta H}}{\beta^2}$ introduced in the sample complexity in contrast to the risk-neutral setting. Finally, notice that for $\beta \approx 0$, the χ^2-divergence admits the expansion $\chi^2(\mathbb{P}_\beta^\pi, \mathbb{P}^\pi) = \beta^2 \mathrm{Var}_{\mathbb{P}^\pi}\left(R_1^\pi|S_1 = s_1\right) + \mathcal{O}(\beta^3)$, which reduces the term in Eq.(9) to the variance term that governs the risk-neutral case.

Algorithm 1 Entropic-BPI

Input: $\beta \neq 0$, $\delta \in (0,1)$, $\varepsilon > 0$.
Initialize counts $n_h^0(\cdot) = 0$ and $\widehat{p}_h^0(\cdot|s,a) = 1/S$.
for $t = 0, 1, 2, \ldots$ **do**
  Terminal init: set $\widetilde{Z}_{H+1}^t(s) = 1$, $\underset{\sim}{Z}_{H+1}^t(s) = 1$, and $G_{H+1}^t(s) = 0$ for all s.
  for $h = H, H-1, \ldots, 1$ **do**
    Compute the bonus $b_h^t(\cdot)$: use (12) if $\beta > 0$, and (18) if $\beta < 0$
    Backup: for all (s,a) compute the optimistic and pessimistic quantities: use (13) if $\beta > 0$, and (19) if $\beta < 0$.
    Greedy action: for all s set

$$\pi_h^{t+1}(s) \in \begin{cases} \arg\max_{a \in \mathcal{A}} \widetilde{U}_h^t(s,a), & \beta > 0, \\ \arg\min_{a \in \mathcal{A}} \widetilde{U}_h^t(s,a), & \beta < 0, \end{cases}$$

    Certificate: Compute $\pi_h^{t+1} G_h^t$: use (14) if $\beta > 0$, and (20) if $\beta < 0$.
  end for
  Stopping test:
  if $\beta > 0$ **and** $(\pi_1^{t+1} G_1^t)(s_1) \leqslant \frac{(e^{\beta\varepsilon}-1)}{e^{\beta\varepsilon}} \widetilde{Z}_1^t(s_1)$ **then**
    stop and output π^{t+1}.
  end if
  if $\beta < 0$ **and** $(\pi_1^{t+1} G_1^t)(s_1) \leqslant (1 - e^{\beta\varepsilon}) \underset{\sim}{Z}_1^t(s_1)$ **then**
    stop and output π^{t+1}.
  end if
  Execute episode $t+1$ with π^{t+1}, update counts and $\widehat{p}_h^{t+1}$.
end for

Algorithm and sample complexity upper bounds. We summarize the elements described above into an algorithm we call ENTROPIC-BPI (Algorithm 1). Using the optimistic proxies in Eq. (13) for $\beta > 0$ and in Eq. (19) for $\beta < 0$, it builds exploratory trajectories until our stopping criterion (Eq.8 for $\beta > 0$ and Eq.(8) for $\beta < 0$) is reached. We prove a sample complexity bound for ENTROPIC-BPI in the following theorem. This complexity bound is valid for both $\beta > 0$ (proof in Appendix B.1) and for $\beta < 0$ (proof in Appendix B.2)

Theorem 4 (sample complexity) *For any $\delta \in [0,1]$ and $\varepsilon > 0$ small enough and for any finite MDP $\mathcal{M}$,* ENTROPIC-BPI *(Algorithm 1) outputs a policy that is (ε, δ)-PAC for best policy identification problem for the entropic risk measure after τ episodes. Moreover, with probability $1 - \delta$:*

$$\tau = \mathcal{O}\left(\frac{e^{2|\beta|\varepsilon}}{\left(e^{|\beta|\varepsilon} - 1\right)^2} \frac{(e^{|\beta| G_{max}(\mathcal{M})} - 1)^2}{e^{|\beta| G_{max}(\mathcal{M})}} SAH \log(\frac{3SAH}{\delta}) \right)$$

The algorithm upper bound matches the lower bound derived in Theorem 3 up to a factor $e^{2|\beta|\varepsilon}$ which is a constant when ε is small enough and comes from having a stopping rule using computable proxies instead of Z_h^π. When $|\beta|$ goes to 0, the upper bound approaches :

$$\tilde{\mathcal{O}}\left(\frac{G_{\max}(\mathcal{M})^2 SAH}{\varepsilon^2} \right) = \tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2} \right)$$

and we recover the optimal sample complexity for the risk-neutral setting (Menard et al., 2021).

Also remark that using the elementary inequality $\log(1+x) \geqslant \frac{x}{2}$ for $|\beta|\varepsilon \in [0,1]$ and using that $\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}} \leqslant e^{|\beta|G_{\max}(\mathcal{M})} - 1 \leqslant e^{|\beta|H} - 1$ we have an upper bound of the order of :

$$\tau = \tilde{\mathcal{O}}\left(\frac{e^{|\beta|H}-1}{\beta^2}\frac{SAH}{\varepsilon^2}\right)$$

This matches the lower bound of Mortensen and Talebi (2025) when mapped to the finite-horizon setting, up to an additional factor H which is unavoidable as it is inherent to the non-stationary finite-horizon setting (with H separate kernels).

Note that the upper bound is stated in terms of $G_{\max}(\mathcal{M})$ rather than H. Since rewards lie in $[0,1]$, $G_{\max}(\mathcal{M})$ can be interpreted as an effective reward horizon, i.e., the maximal cumulative reward that can be accrued along a trajectory. This choice is natural for the entropic risk criterion, whose difficulty comes from the exponential amplification of accumulated rewards; using H may overestimate this effect in problems where rewards are sparse or concentrated near the end of the episode. Finally, our algorithm does not require prior knowledge of $G_{\max}(\mathcal{M})$.

Experiments. To illustrate the gains in sample complexity, we propose a simple 8-state MDP and compare ENTROPIC-BPI with regret algorithms in the literature. The results are discussed in Appendix E

5. Proof of Theorem 4

We present the main ideas of the proof for $\beta > 0$. The full proof is given in Appendix B. We first control the concentration events (Lemma 6) and work on the good event $\mathcal{E}^+$ for $\beta > 0$, which holds with probability at least $1-\delta$. As explained in paragraph Stopping rule, when the algorithm stops, it outputs by design a policy π^τ that is ε-optimal with high probability $1-\delta$, this proves the first statement of Theorem 4.

To upper bound the stopping time, we first show that $\widetilde{U}_h^t$ and $\underset{\sim}{U}_h^t$ are indeed optimistic and pessimistic, respectively, for the exponential transform of the value functions $U_h^\star$ (Lemma 9). Then, following a similar approach to (Menard et al., 2021; Dann et al., 2017), we bound the width certificate by a computable recursive upper bound, which serves as the stopping rule for our algorithm (6). Lemma 12 shows that, with probability at least $1-\delta$, this width certificate upper bounds the optimality gap. Since the stopping condition is not met for episodes $t = \{1,\dots,\tau\}$, we have:

$$\tau\frac{e^{\varepsilon\beta}-1}{e^{\varepsilon\beta}} \leqslant \sum_{t=1}^{\tau}\frac{\pi_1 G_1^t(s_1)}{\widetilde{Z}_1^t(s_1)}$$

We bound the right-hand side for episode t by replacing the empirical transition probabilities with the true ones. We then unroll the resulting recursive inequality for $\pi_{h+1}^t G_h^t$ under the true dynamics (see B.1.3 for details):

$$\frac{(\pi_1 G_1^t)(s_1)}{\widetilde{Z}_1^t(s_1)} \leqslant e^{13}\mathbb{E}^{\pi^{t+1}}\left[\sum_{h=1}^{H}\exp\left(\beta\sum_{i=1}^{h}r_i(s_i,a_i)\right)\left(36\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\alpha^\star\left(n_h^t(s_h,a_h)\wedge 1\right)}\right.\right.$$
$$\left.\left. + 81He^{\beta(H-h)}\alpha\left(n_h^t(s_h,a_h)\wedge 1\right)\right)\middle| s_1\right]$$

We bound the first term using Cauchy-Schwartz inequality:

$$\mathbb{E}^{\pi^{t+1}}\left[\sum_{h=1}^{H}\exp\Big(\beta\sum_{i=1}^{h}r_i(s_i,a_i)\Big)\left(\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)}\right)\middle| s_1\right]$$
$$\leqslant\sqrt{\mathbb{E}^{\pi^{t+1}}\Big[\exp\Big(2\beta\sum_{i=1}^{h}r_i(s_i,a_i)\Big)\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\Big]}\sqrt{\mathbb{E}^{\pi^{t+1}}\Big[\frac{\alpha^\star(n_h^t(s,a))\big)}{n_h^t(s,a)}\Big]}$$

Using lemma 25 and lemma 26 we bound the first term on the right-hand side as:

$$\sum_{h=1}^{H}\mathbb{E}^{\pi^{t+1}}\left[\exp\Big(2\beta\sum_{i=1}^{h}r_i(s_i,a_i)\Big)\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\right]=\frac{\mathrm{Var}\left(e^{\beta R_1^{\pi^{t+1}}}(S_1)\middle|S_1=s_1\right)}{\left(\mathbb{E}\left[e^{\beta R_1^{\pi^{t+1}}(S_1)}\middle|S_1=s_1\right]\right)^2}\leqslant\frac{\left(e^{\beta G_{\max}(\mathcal{M})}-1\right)^2}{e^{\beta G_{\max}(\mathcal{M})}}$$

For the second term, we bound it loosely by directly upper bounding the per-step reward by 1:

$$\mathbb{E}^{\pi^{t+1}}\left[\sum_{h=1}^{H}\exp\Big(\beta\sum_{i=1}^{h}r_i(s_i,a_i)\Big)He^{\beta(H-h)}\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)\right)\middle|s_1\right]\leqslant He^{\beta H}\mathbb{E}\Big[\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\Big]$$

We sum over all episodes $t=1,...,\tau$. Then using a standard counting argument we have:

$$\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\mathbb{E}^{\pi^{t+1}}\Big[\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\Big]\leqslant 3SAH\alpha^\star(\tau-1,\delta)\log(\tau+1)$$

And we have a similar bound for α, Hence:

$$\tau\frac{e^{\beta\varepsilon}-1}{e^{\beta\varepsilon}}\leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\tau SAH\alpha^\star(\tau-1,\delta)\log(\tau+1)}$$
$$+\,84e^{13}e^{\beta H}SAH\alpha(t-1,\delta)\log(t+1)$$

Solving this inequality using lemma 31, we find the exact upper bound on τ with probability $1-\delta$.

6. Regret bounds

Although the main focus of this work is best-policy identification, the techniques developed above also have implications for regret minimization. In particular, the certificate analysis for Entropic-BPI can be used to control the regret of a non-stopping variant of the algorithm.

We emphasize that this is not a direct reduction from BPI: BPI evaluates only the final policy, whereas regret evaluates the full learning trajectory. Nevertheless, both problems are driven by the need to identify rare transitions whose contribution is exponentially amplified by the entropic criterion.

We assume in this section that $\beta>0$ but the same analysis follows for $\beta<0$. We consider the non-stopping variant of algorithm 1. At each episode, the algorithm outputs the optimistic model from the current empirical counts, plays the greedy policy, updates the counts and continues indefinitely.

We define the cumulative regret after K episodes as:

$$R(K) = \sum_{t=0}^{K-1} (V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1)) \tag{10}$$

The key observation is that the same certificate G^t used in the stopping rule controls each instantaneous regret term through the normalized ratio $\pi_1 G_1(s_1)/Z_1^{\pi_1^t}(s_1)$. Controlling this normalized ratio is what allows us to recover the sharper exponential dependence

Theorem 5 *On the same high-probability event as theorem 4, the non-stopping variant of Algorithm 1 satisfies, for every $K \geqslant 2$,*

$$R(K) = \mathcal{O}\left(\frac{1}{|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}} KSAH \log\left(\frac{SAHK}{\delta}\right)} + \frac{e^{|\beta|H}}{\beta} S^2AH\log^2\left(\frac{SAHK}{\delta}\right)\right).$$

In particular when $G_{\max}(\mathcal{M})$ is of the order of H, the exponential dependency becomes $\frac{e^{\beta H/2}}{\beta}$, improving over existing regret bounds whose leading exponential dependence is $\frac{e^{\beta H}}{\beta}$.

The proof follows the sample-complexity analysis, but instead of stopping when the certificate becomes smaller than the desired accuracy threshold, we sum the normalized certificates over the first K episodes. The variance sensitive part is again controlled through the exponential Bellman variance identity, which yields the factor $\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}$. The cumulative exploration terms are then bounded using the same pseudo-count argument as in the stopping-time analysis.

We also evaluate the non-stopping variant of ENTROPIC-BPI on the toy MDP described in Appendix E. The comparison includes several risk-sensitive regret algorithms from the literature. Across the tested values of β, our method achieves smaller cumulative regret, and the improvement becomes more noticeable as the risk sensitivity increases, supporting the sharper exponential dependence in Theorem 5.

Regret lower bound The lower bound construction of section 3 can be adapted to get the lower bound for the regret setting. Recall that the hard family contains $\Theta(SAH)$ possible hard triplets $u = (h^\star, \ell^\star, a^\star)$. In instance M_u, only the triplet u has the favorable transition probability p_+, while all other triplets have probability p_-. Thus, before the learner gathers enough information to identify u, it incurs regret whenever it fails to visit the correct triplet.

The regret lower bound is obtained by choosing the value gap ε as a function of the time horizon K. The information gained from one visit to the special triplet is of order : $\frac{\beta^2\varepsilon^2}{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}}$. Since there are $\Theta(SAH)$ possible hard triplets and only K episodes, the learner cannot identify the correct triplet when ε is the order of $\frac{1}{|\beta|}\sqrt{\frac{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}SAH}{K}}$. Consequently, the expected regret satisfies:

$$\mathbb{E}[R(K)] \geqslant \frac{c_0}{|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}SAHK}$$

This regret lower bound improves on existing lower bounds in the literature. Combining the regret upper bound with the corresponding lower-bound argument shows that the non-stopping optimistic variant of Entropic-BPI is minimax-optimal in its leading $\sqrt{K}$ term, up to logarithmic factors

and lower-order additive terms. In particular, in worst-case instances where $G_{\max}(\mathcal{M}) \asymp H$, the leading exponential factor is $e^{|\beta|H/2}/|\beta|$. Thus, at the level of the leading regret term, the analysis closes the gap between previous $e^{|\beta|H}/|\beta|$-type upper bounds and the $e^{|\beta|H/2}/|\beta|$ lower-bound dependence.

7. Discussions and Conclusions

We provide a new approach to entropic best arm identification that resolves a known suboptimality gap. Our approach builds a successful optimism-driven framework for the forward model, relying on a tight control of the variance of the estimators of the entropic value function, and on a specifically tailored stopping rule.

Indeed, the dependence of the sample complexity on the horizon remains exponential, indicating one more time that learning exponential utilities in MDPs is a significantly harder problem than the standard expected return due to the focus on tail (rare) events. However, we show that the real MDP-dependent term of interest in the exponential is the maximal return, which could be constant in some sparse reward problems, making the problem more amenable. Such problem-specific investigations could be an avenue for future work. More broadly, this exponential dependence should be interpreted as the intrinsic statistical price of entropic risk sensitivity rather than as evidence that the criterion is unsuitable for learning. Indeed, the entropic risk measure has an important structural advantage: it preserves an exact Bellman recursion on the original MDP state space and is essentially characterized by dynamic consistency among utility-based objectives (Marthe et al., 2023). By contrast, other tail-risk criteria such as CVaR can enjoy favorable statistical rates, including near-minimax regret guarantees with polynomial dependence on the tail parameter (Wang et al., 2023). However, these guarantees rely on additional assumptions on the return distributions, and direct optimization of CVaR typically requires state augmentation and optimization over an additional risk variable, often handled through discretization which may lead to inefficient runtime.

Recently, Marthe et al. (2025) showed that there is a fundamental connection between the entropic risk measure and more practically-used metrics such as the (conditional) Value at Risk. Our forward-model approach could be combined with such optimization improvements to propose new algorithms for (C)VaR optimization in RL.

Acknowledgments

C. Vernade is funded by the Deutsche Forschungsgemeinschaft (DFG) under both the project 468806714 of the Emmy Noether Programme and under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. CV also gratefully acknowledges funding from the European Union (ERC, ConSequentIAL, 101165883). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them). CV also thanks the international Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

Mohammad Gheshlaghi Azar, Remi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference*

on Machine Learning (ICML-12), ICML '12, pages 1263–1270, New York, NY, USA, July 2012. Omnipress. ISBN 978-1-4503-1285-1.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/azar17a.html.

V. S. Borkar and S. P. Meyn. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. URL https://doi.org/10.1137/S0363012997331639.

Arthur Charpentier, Romuald Élie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, 62(1):425–462, 2023. doi: https://doi.org/10.1007/s10614-021-10119-4.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2 edition, 2006. ISBN 9780471241959. URL https://doi.org/10.1002/047174882X.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/309fee4e541e51de2e41f21bebb342aa-Paper.pdf.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 578–598. PMLR, 16–19 Mar 2021. URL https://proceedings.mlr.press/v132/domingues21a.html.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems*, volume 33, pages 22384–22395. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fdc42b6b0ee16a2f866281508ef56730-Paper.pdf.

Yingjie Fei, Zhuoran Yang, Yudong Chen, and Zhaoran Wang. Exponential bellman equation and improved regret bounds for risk-sensitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20436–20446. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ab6439fa2daf0246f92eea433bca5ac4-Paper.pdf.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the Seventh Annual Conference on Computational Learning Theory*, COLT '94, page 88–97, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916557. URL https://doi.org/10.1145/180139.181019.

Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972. doi: https://doi.org/10.1287/mnsc.18.7.356.

Xiaoyan Hu and Ho-Fung Leung. A tighter problem-dependent regret bound for risk-sensitive reinforcement learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5411–5437. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/hu23b.html.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL http://jmlr.org/papers/v17/kaufman16a.html.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 2021. URL https://proceedings.mlr.press/v132/kaufmann21a.html.

Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49:193–208, 2002. URL https://doi.org/10.1023/A:1017932429737.

Hao Liang and Zhi-Quan Luo. Bridging distributional and risk-sensitive reinforcement learning with provable regret bounds. *Journal of Machine Learning Research*, 25(221):1–56, 2024. URL http://jmlr.org/papers/v25/22-1253.html.

Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in markov decision processes. In *Advances in Neural Information Processing Systems*, volume 36, pages 56488–56507. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b0a34e3c64f7e842f20ec10479c32b35-Paper-Conference.pdf.

Alexandre Marthe, Samuel Bounan, Aurélien Garivier, and Claire Vernade. Efficient risk-sensitive planning via entropic risk measures. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2502.20423. URL https://arxiv.org/abs/2502.20423.

Pierre Menard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In

Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pages 7599–7608. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/menard21a.html.

Oliver Mortensen and Mohammad Sadegh Talebi. Entropic risk optimization in discounted mdps: Sample complexity bounds with a generative model. *arXiv preprint arXiv:2506.00286*, 2025. URL https://arxiv.org/abs/2506.00286.

Athanasios S. Polydoros and Lazaros Nalpantidis. Survey of model-based reinforcement learning: Applications on robotics. *Journal of Intelligent & Robotic Systems*, 86(2):153–173, 2017. doi: https://doi.org/10.1007/s10846-017-0468-y.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Remi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5528–5536. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rowland19a.html.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(84):2413–2444, 2009. URL http://jmlr.org/papers/v10/strehl09a.html.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 07–09 Apr 2018. URL https://proceedings.mlr.press/v83/talebi18a.html.

Kaiwen Wang, Nathan Kallus, and Wen Sun. Near-minimax-optimal risk-sensitive reinforcement learning with CVaR. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35864–35907. PMLR, 23–29 Jul 2023.

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

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learningvia reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, pages 15198–15207. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ad71c82b22f4f65b9398f76d8be4c615-Paper.pdf.

Appendix

Contents of Appendix

Appendix A. Concentration events

Following the ideas of (Menard et al., 2021) we define the following quantities:

$$\begin{aligned}\alpha(n,\delta) &= \log(3SAH/\delta) + S\log(8e(n+1)) \\ \alpha^{\text{cnt}}(\delta) &= \log(3SAH/\delta) \quad \text{and} \\ \alpha^{\star}(n,\delta) &= \log(3SAH/\delta) + \log(8e(n+1))\end{aligned} \tag{11}$$

We also define the KL-divergence concentration event as:

$$\mathcal{E}_{KL} = \left\{\forall t \in \mathbb{N}, \forall h \in \{1,...,H\}, \forall (s,a) \in \mathcal{S}\times\mathcal{A} : D_{KL}(\widehat{p}_h^t(s,a), p_h(s,a)) \leqslant \alpha(n_h^t(s,a),\delta)\right\}$$

and the Bernoulli concentration event for a series of function $(f_h)_{h\in[H+1]}$ in $[0,b]$:

$$\mathcal{E}_f = \Bigg\{\forall t \in \mathbb{N}, \forall h \in \{1,...,H\}, \forall (s,a) \in \mathcal{S}\times\mathcal{A} : \\ \left|(\widehat{p}_h^t - p_h) f_{h+1}(s,a)\right| < \sqrt{2\,\mathrm{Var}_{p_h}(f_{h+1})(s,a)\frac{\alpha^{\star}(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3b\frac{\alpha^{\star}(n_h^t(s,a),\delta)}{n_h^t(s,a)}\Bigg\}$$

And the counts concentration event:

$$\mathcal{E}^{\text{cnt}} = \left\{\forall t \in \mathbb{N}, \forall h \in \{1,...,H\}, \forall (s,a) \in \mathcal{S}\times\mathcal{A} : n_h^t(s,a) \geqslant \frac{1}{2}\bar{n}_h^t(s,a) - \alpha^{\text{cnt}}(\delta)\right\}$$

Using this, we define the good event for our algorithm analysis for $\beta > 0$ and $\beta < 0$: for $\beta > 0$

$$\mathcal{E}^{+} = \mathcal{E}_{KL} \cap \mathcal{E}_{V*} \cap \mathcal{E}^{\text{cnt}}$$

And for $\beta < 0$ we have almost the same thing but the definition of $\mathring{V}$ and the range of the functions changes:

$$\mathcal{E}^{-} = \mathcal{E}_{KL} \cap \mathcal{E}_{V*} \cap \mathcal{E}^{\text{cnt}}$$

Lemma 6 *It holds that :*

$$\mathbb{P}(\mathcal{E}_{KL}) \geqslant 1-\frac{\delta}{3}, \qquad \mathbb{P}(\mathcal{E}^{cnt}) \geqslant 1-\frac{\delta}{3}, \qquad \textit{and} \qquad \textit{for any } f \quad \mathbb{P}(\mathcal{E}_f) \geqslant 1-\frac{\delta}{3}$$

Consequently,

$$\mathbb{P}(\mathcal{E}^{+}) \geqslant 1-\delta \quad \textit{and} \quad \mathbb{P}(\mathcal{E}^{-}) \geqslant 1-\delta$$

Proof

The KL concentration event:

For (h,s,a) fixed, we apply lemma F.1 with confidence level $\delta_{h,s,a} = \frac{\delta}{3SAH}$ and then do a union bound over h,s,a to get a concentration inequality that holds uniformly.

The Bernstein concentration event:

Let $(f_h)_{h\in[H+1]}$ be a sequence of function. Fix (h,s,a) and let $(\mathcal{F}_\tau)_{\tau\geqslant 0}$ be the history filtration. Define

$$w_\tau = \mathbf{1}\{(H_\tau,S_\tau,A_\tau) = (h,s,a)\}, \qquad Y_\tau = f_{h+1}(S_{\tau+1}) - \mathbb{E}[f_{h+1}(S_{\tau+1}) \mid \mathcal{F}_{\tau-1}]$$

Then (w_τ) is predictable, $\mathbb{E}[Y_\tau \mid \mathcal{F}_{\tau-1}] = 0$, $|Y_\tau| \leqslant H$, and

$$\mathbb{E}[Y_\tau^2 \mid \mathcal{F}_{\tau-1}] = \mathrm{Var}_{p_h}(f_{h+1})(s,a) \quad \text{on } \{w_\tau = 1\}$$

Let

$$S_t = \sum_{\tau=1}^{t} w_\tau Y_\tau, \quad V_t = \sum_{\tau=1}^{t} w_\tau^2 \mathbb{E}[Y_\tau^2 \mid \mathcal{F}_{\tau-1}], \quad W_t = \sum_{\tau=1}^{t} w_\tau = n_h^t(s,a)$$

Then $V_t = W_t \,\mathrm{Var}_{p_h}(f_{h+1})(s,a)$ and, for $W_t \geqslant 1$,

$$(\widehat{p}_h^t - p_h) f_{h+1}(s,a) = \frac{S_t}{W_t}$$

Applying lemma 24 with b and confidence $\delta_{h,s,a}$ yields (simultaneously for all t)

$$\left|(\widehat{p}_h^t - p_h) f_{h+1}(s,a)\right| \leqslant \sqrt{2\,\mathrm{Var}_{p_h}(f_{h+1})(s,a)\frac{\alpha^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3B\frac{\alpha^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

where $\alpha^\star(n,\delta) = \log\left(\frac{4e(2n+1)}{\delta_{h,s,a}}\right)$ (using $n_h^t(s,a) \leqslant t$ and monotonicity of the log term). Finally choose $\delta_{h,s,a} = \frac{\delta}{SAH}$ and we apply a union bound over (h,s,a) to get the result for any state-action pair

The counts concentration event:

The proof follows from lemma 22 applied to the Bernoulli random variable $\mathbb{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}$ for $\delta_{h,s,a} = \frac{\delta}{SAH}$ and then doing a union bound over h, s, a ■

Appendix B. Algorithm analysis

Here we provide a detailed analysis of our algorithm. Our method is a UCB-style algorithm that plans over a KL confidence region, following the approach of Menard et al. (2021) for the risk-neutral objective. At each step t and stage h, we construct a confidence set around the true transition kernel:

$$\mathcal{C}_h^t(s,a) \triangleq \left\{ q \in \Sigma_S : \mathrm{KL}\left(\widehat{p}_h^t(s,a), q(s,a)\right) \leqslant \frac{\alpha(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$$

The algorithm then acts optimistically by selecting, among all transition models q such that $q(.|s,a) \in \mathcal{C}_h^t(s,a)$, the one that yields the highest value function, and plans accordingly.

For the entropic risk measure, we follow the same principle, but carry out optimistic planning in the exponential (log-moment-generating) space induced by the entropic criterion. As noted by Fei et al. (2021), working in this exponential space allows us to exploit a Bellman-type recursion that is generally lost if one applies Lipschitz arguments directly in the original value space. This means that the upper and lower confidence bounds on the optimal exponential transformation of the

state-value function U^* and value function Z^* for $\beta > 0$ are given by:

$$\overline{U}_h^t(s,a) \triangleq e^{\beta r_h(s,a)} \max_{\overline{p}_h \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{Z}_{h+1}^t(s,a) \qquad \underline{U}_h^t(s,a) \triangleq e^{\beta r_h(s,a)} \min_{\underline{p}_h \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{Z}_{h+1}^t)(s,a)$$

$$\overline{Z}_h^t(s) \triangleq \max_a \overline{U}_h^t(s,a) \qquad \underline{Z}_h^t(s) \triangleq \max_a \underline{U}_h^t(s,a)$$

$$\overline{Z}_{H+1}^t(s) \triangleq 0 \qquad \underline{Z}_{H+1}^t(s) \triangleq 0$$

$$\overline{p}_h^t(s,a) \in \arg\max_{\overline{p} \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{Z}_{h+1}^t(s,a) \qquad \underline{p}_h^t(s,a) \in \arg\min_{\underline{p} \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{Z}_{h+1}^t(s,a)$$

$$\overline{\pi}_h^t(s,a) \in \arg\max_{a \in \mathcal{A}} \overline{U}_h^t(s,a) \qquad \underline{\pi}_h^t(s,a) \in \arg\max_{a \in \mathcal{A}} \underline{U}_h^t(s,a).$$

For $\beta < 0$, $\overline{U}$ will correspond to the pessimistic $\underline{Q}$ via the log-transformation. As such, maximizing $\overline{Q}$ to define the policy $\overline{\pi}$ is equivalent to minimizing $\underline{U}$. Similarly, finding the best action at each stage to define $\overline{Z}$ and $\underline{Z}$ corresponds to minimizing $\overline{U}$ and $\underline{U}$ respectively:

$$\overline{U}_h^t(s,a) \triangleq e^{\beta r_h(s,a)} \min_{\overline{p}_h \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{Z}_{h+1}^t(s,a) \qquad \underline{U}_h^t(s,a) \triangleq e^{\beta r_h(s,a)} \max_{\underline{p}_h \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{Z}_{h+1}^t(s,a)$$

$$\overline{Z}_h^t(s) \triangleq \min_a \overline{U}_h^t(s,a) \qquad \underline{Z}_h^t(s) \triangleq \min_a \underline{U}_h^t(s,a)$$

$$\overline{Z}_{H+1}^t(s) \triangleq 0 \qquad \underline{Z}_{H+1}^t(s) \triangleq 0$$

$$\overline{p}_h^t(s,a) \in \arg\min_{\overline{p} \in \mathcal{C}_h^t(s,a)} \overline{p}_h \overline{Z}_{h+1}^t(s,a) \qquad \underline{p}_h^t(s,a) \in \arg\max_{\underline{p} \in \mathcal{C}_h^t(s,a)} \underline{p}_h \underline{Z}_{h+1}^t(s,a)$$

$$\overline{\pi}_h^t(s,a) \in \arg\min_{a \in \mathcal{A}} \overline{U}_h^t(s,a) \qquad \underline{\pi}_h^t(s,a) \in \arg\max_{a \in \mathcal{A}} \underline{U}_h^t(s,a).$$

The KL confidence sets $C_h^t(s,a)$ are introduced solely to motivate an optimistic model interpretation. We instead build computable optimistic and pessimistic expressions in the empirical MDP by choosing the radius α so that the true transition kernel belongs to $\mathcal{C}_h^t(s,a)$ in the same style as (Menard et al., 2021). We then prove the corresponding optimism lemma, bound the certificate width, and derive the sample complexity. We treat the cases $\beta > 0$ and $\beta < 0$ separately. We first restate the theorem in more detail

Theorem 4 (sample complexity) *For any $\delta \in [0,1]$ and $\varepsilon \in]0, \frac{2}{|\beta|HS}]$ and for any finite MDP $\mathcal{M}$,* ENTROPIC-BPI *(Algorithm 1) outputs a policy that is (ε,δ)-PAC for best policy identification problem for the entropic risk measure after τ episodes. Moreover, with probability $1-\delta$:*

$$\tau \leqslant \frac{e^{2\max\{0,\beta\}\varepsilon}}{\left(e^{|\beta|\varepsilon}-1\right)^2} \frac{\left(e^{|\beta|G_{max}(\mathcal{M})}-1\right)^2}{e^{|\beta|G_{max}(\mathcal{M})}} SAH \log\left(\frac{3SAH}{\delta}\right) C_2^2$$

Where $C_2 = 2765e^{22}\log\left(4e^{17}\frac{(S+1)(H+1)e^{|\beta|H}SAH^2}{e^{\beta\varepsilon}-1}\right)$. *In particular, hiding constants and log terms:*

$$\tau = \tilde{\mathcal{O}}\left(\frac{e^{2\max\{0,\beta\}\varepsilon}}{\left(e^{|\beta|\varepsilon}-1\right)^2} \frac{\left(e^{|\beta|G_{max}(\mathcal{M})}-1\right)^2}{e^{|\beta|G_{max}(\mathcal{M})}} SAH\right)$$

B.1. Case $\beta > 0$

. We start by building optimistic and pessimistic functions for the state-value function

B.1.1. CONFIDENCE BOUNDS

Let us start with a concentration inequality:

Lemma 8 *On the good event $\mathcal{E}^+$ we have:*

$$|(p_h - \widehat{p}_h^t) Z_{h+1}^\star(s,a)| \leqslant 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + 5e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + 4He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a)$$

Proof On the good event $\mathcal{E}^+$ we have:

$$|(p_h - \widehat{p}_h^t) Z_{h+1}^\star(s,a)| \leqslant \sqrt{2\,\mathrm{Var}_{p_h}(Z_{h+1}^\star)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 3e^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}$$

We apply lemma 27 and lemma 28 successively to transport the variance of $Z_{h+1}^\star$ under p_h to the computable variance of $\widetilde{Z}_{h+1}^t$ under $\widehat{p}_h^t$

$$\begin{aligned}\mathrm{Var}_{p_h}(Z_{h+1}^\star) &\leqslant 2\,\mathrm{Var}_{\widehat{p}_h^t}(Z_{h+1}^\star)(s,a) + 4e^{2\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\ &\leqslant 4\,\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)(s,a) + 4e^{\beta(H-h)}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star)(s,a) + 4e^{2\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\end{aligned}$$

Hence, using the inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ and then $\sqrt{ab} \leqslant a\eta + \frac{b}{\eta}$ for $\eta = H$ and using that $\alpha^\star(n,\delta) \leqslant \alpha(n,\delta)$:

$$\begin{aligned}\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^\star)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant 2\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_h^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \sqrt{4e^{\beta(H-h)}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\ &\quad + 2e^{\beta(H-h)}\sqrt{\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\ &\leqslant 2\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_h^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\ &\quad + \frac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star)(s,a) + 2e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\end{aligned}$$

Hence, by plugging this upper bound and using again that $\alpha^\star(n,\delta) \leqslant \alpha(n,\delta)$ we obtain:

$$|(p_h - \widehat{p}_h^t) Z_{h+1}^\star| \leqslant 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a)$$

■

Denote the bonus term:

$$b_h^t(s,a) = 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \tag{12}$$

Now, following Azar et al. (2017), Zanette and Brunskill (2019) and Menard et al. (2021) we define optimistic and pessimistic state-value function on the exponential transform of $Q_h^\star$ which is denoted by $U_h^\star$:

$$\begin{aligned}
\widetilde{U}_h^t(s,a) &= \min\left\{e^{\beta(H-h-1)}, e^{\beta r_h(s,a)}\left[\widehat{p}_h^t\widetilde{Z}_h^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]\right\} \\
\widetilde{Z}_h^t(s) &= \max_{a\in\mathcal{A}}\widetilde{U}_h^t(s,a), \qquad \widetilde{Z}_{H+1}^t(s) = 1 \\
\underset{\sim}{U}_h^t(s,a) &= \max\left\{1, e^{\beta r_h(s,a)}\left[\widehat{p}_h^t\underset{\sim}{Z}_h^t(s,a) - b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]\right\} \\
\underset{\sim}{Z}_h^t(s) &= \max_{a\in\mathcal{A}}\underset{\sim}{U}_h^t(s,a), \qquad \underset{\sim}{Z}_{H+1}^t(s) = 1
\end{aligned} \tag{13}$$

And we consider the greedy policy:

$$\pi_h^{t+1}(s) = \arg\max_{a\in\mathcal{A}}\widetilde{U}_h^t(s,a)$$

Now let us prove the optimism lemma:

Lemma 9 *With high probability $1-\delta$ we have:*

$$\underset{\sim}{U}_h^t(s,a) \leqslant U_h^\star(s,a) \leqslant \widetilde{U}_h^t(s,a)$$

and

$$\underset{\sim}{Z}_h^t(s) \leqslant Z_h^\star(s) \leqslant \widetilde{Z}_h^t(s)$$

Proof We proceed by induction over h. For $h = H+1$ the result is trivially upper bounding and (resp. lower bounding) $U_h^\star$ by $e^{\beta(H-h)}$ and 1.
Assume the inequality holds for $h' > h$. Fix (s,a) and assume $\widetilde{U}_h^t(s,a) < H$ since otherwise the inequality is trivial, we have that:

$$\begin{aligned}
\widetilde{U}_h^t(s,a) - U_h^\star(s,a) &= e^{\beta r_h(s,a)}\left[\widehat{p}_h^t\widetilde{Z}_{h+1}^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) - p_h Z_{h+1}^\star(s,a)\right] \\
&= e^{\beta r_h(s,a)}\left[\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t(s,a) - Z_{h+1}^\star(s,a)\Big) + \Big(\widehat{p}_h^t - p_h\Big)Z_{h+1}^\star(s,a)\right. \\
&\qquad \left. + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]
\end{aligned}$$

But we know by Bernstein inequality that:

$$\Big(\widehat{p}_h^t - p_h\Big) Z_h^\star(s,a) \geqslant -b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a)$$

Hence:

$$\widetilde{U}_h^t(s,a) - U_h^\star(s,a) \geqslant e^{\beta r_h(s,a)}\left[\Big(1-\frac{1}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t(s,a) - Z_{h+1}^\star(s,a)\Big) + \frac{1}{H}\widehat{p}_h^t\Big(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right] \geqslant 0$$

Where we used the induction hypothesis. We prove the pessimistic property in the same way ■

B.1.2. STOPPING RULE

We define the width certificate for the algorithm for the case $\beta > 0$:

$$G_h^t(s,a) = \min\left\{e^{\beta(H-h)}, e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t\pi^{t+1}G_{h+1}^t(s)\Big]\right\} \tag{14}$$

Lemma 10 establishes the validity of this stopping rule by showing that, with high probability, it bounds the certificate width:

Lemma 10 *On the good event $\mathcal{E}^+$, for all t and all h,*

$$Z_h^\star(s) - Z_h^{\pi^{t+1}}(s) \leqslant \pi_h^{t+1}G_h^t(s) \qquad \forall s \in \mathcal{S}$$

In particular, at the initial state s_1, $Z_1^\star(s_1) - Z_1^{\pi^{t+1}}(s_1) \leqslant \pi_1^{t+1}G_1^t(s_1)$

We prove the lemma 10 in this section : We define the auxiliary variable $\mathring{Z}_h^t$. Setting $\mathring{Z}_{H+1}^t \equiv 1$, we recurse backward for $h = H \dots 1$:

$$\begin{aligned}
\mathring{U}_{h,\text{pes}}^t &= \max\left\{1, e^{\beta r_h}\left[\widehat{p}_h^t\mathring{Z}_{h+1}^t - b_h^t - \tfrac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t)\right]\right\}\\
\mathring{U}_h^t &= \min\left\{e^{\beta r_h}(p_h\mathring{Z}_{h+1}^t), \mathring{U}_{h,\text{pes}}^t\right\}\\
\mathring{Z}_h^t(s) &= \mathring{U}_h^t\big(s, \pi_h^{t+1}(s)\big)
\end{aligned}$$

Because $\underset{\sim}{Z}^t$ is pessimistic against $Z^\star$, we cannot directly compare $\underset{\sim}{Z}^t$ to $Z^{\pi^{t+1}}$. Intuitively, $\mathring{Z}^t$ satisfies the exponential Bellman recursion under the true kernel p_h, while being clipped by a pessimistic empirical backup; hence it serves as a worst-case lower bound for both $\underset{\sim}{Z}^t$ and $Z^{\pi^{t+1}}$ as shows the next lemma:

Lemma 11 *For all (h,s,a):*

$$\mathring{U}_h^t(s,a) \leqslant \min\Big(\underset{\sim}{U}_h^t(s,a), U_h^{\pi_h^{t+1}}(s,a)\Big)$$

and

$$\mathring{Z}_h^t(s) \leqslant \min\Big(\underset{\sim}{Z}_h^t(s), Z_h^{\pi_h^{t+1}}(s,a)\Big)$$

Proof We proceed by backward induction. For $h = H+1$, all values are equal to 1 so the inequalities hold. Assume that for some $h \leqslant H$ we have for all (s,a):

$$\mathring{U}^t_{h+1}(s,a) \leqslant \min\Big(\underset{\sim}{U}{}^t_{h+1}(s,a), U_h^{\pi^{t+1}_{h+1}}(s,a)\Big)$$

and

$$\mathring{Z}^t_{h+1}(s) \leqslant \min\Big(\underset{\sim}{Z}{}^t_{h+1}(s), Z_h^{\pi^{t+1}_{h+1}}(s,a)\Big)$$

we have by construction:

$$\mathring{U}^t_h(s,a) \leqslant \mathring{U}^t_{h,\text{true}}(s,a) = e^{\beta r_h(s,a)}(p_h \mathring{Z}^t_{h+1})(s,a) \leqslant e^{\beta r_h(s,a)}(p_h Z^{\pi^{t+1}}_{h+1})(s,a) = U_h^{\pi^{t+1}_{h+1}}(s,a)$$

Where we used the induction hypothesis and the monotonicity of the exponential Bellman operator. Again by construction;

$$\mathring{U}^t_h(s,a) \leqslant \mathring{U}^t_{h,\text{pes}}(s,a)$$

But since we have:

$$\begin{aligned}
\underset{\sim}{U}{}^t_h(s,a) - \mathring{U}^t_{h,\text{pes}}(s,a) &= e^{\beta r_h(s,a)}\Big[\Big(\widehat{p}^t_h \underset{\sim}{Z}{}^t_{h+1}(s,a) - b^t_h(s,a) - \frac{1}{H}\widehat{p}^t_h\Big(\widetilde{Z}^t_{h+1} - \underset{\sim}{Z}{}^t_{h+1}\Big)(s,a)\Big) \\
&\quad - \Big(\widehat{p}^t_h \mathring{Z}^t_{h+1}(s,a) - b^t_h(s,a) - \frac{1}{H}\widehat{p}^t_h\Big(\widetilde{Z}^t_{h+1} - \mathring{Z}^t_{h+1}\Big)(s,a)\Big)\Big] \\
&= e^{\beta r_h(s,a)}\Big(1 - \frac{1}{H}\Big)\Big[\widehat{p}^t_h\big(\underset{\sim}{Z}{}^t_{h+1} - \mathring{Z}^t_{h+1}\big)\Big] \\
&\geqslant 0
\end{aligned}$$

Where we applied the induction hypothesis, we conclude then:

$$\mathring{U}^t_h(s,a) \leqslant \underset{\sim}{U}{}^t_h(s,a)$$

The bound on V follows immediately and we conclude the recursion ■

On the good event $\mathcal{E}^+$ we have:

Lemma 12 *On the good event $\mathcal{E}^+$*

$$\widetilde{U}^t_h(s,a) - \mathring{U}^t_h(s,a) \leqslant e^{\beta r^t_h(s,a)}\Big[3b^t_h(s,a) + \Big(1 + \frac{3}{H}\Big)\widehat{p}^t_h\Big(\widetilde{Z}^t_{h+1} - \mathring{Z}^t_{h+1}\Big)\Big]$$

Proof Fix a state-action pair (s,a) and $h \in \{1,...,H\}$, we consider two cases:
First case: $\mathring{U}^t_h(s,a) = \mathring{U}^t_{h,\text{true}}(s,a)$ we then have:

$$\widetilde{U}^t_h(s,a) - \mathring{U}^t_h(s,a) \leqslant e^{\beta r^t_h(s,a)}\Big[b^t_h(s,a) + \frac{1}{H}\widehat{p}^t_h\Big(\widetilde{Z}^t_{h+1} - \underset{\sim}{Z}{}^t_{h+1}\Big)(s,a) + \widehat{p}^t_h \widetilde{Z}^t_{h+1}(s,a) - p_h \mathring{Z}^t_{h+1}(s,a)\Big]$$

The last term can be written as:

$$\widehat{p}^t_h \widetilde{Z}^t_{h+1}(s,a) - p_h \mathring{Z}^t_{h+1}(s,a) = \widehat{p}^t_h\Big(\widetilde{Z}^t_{h+1} - \mathring{Z}^t_{h+1}\Big)(s,a) + \big(\widehat{p}^t_h - p_h\big)Z^\star_{h+1} + \big(p_h - \widehat{p}^t_h\big)\big(Z^\star_{h+1} - \mathring{Z}^t_{h+1}\big)$$

For the second term, by lemma 8:

$$|\big(p_h - \widehat{p}_h^t\big) Z_{h+1}^\star(s,a)| \leqslant b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a) \leqslant b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)(s,a)$$

For the third term, by the KL-Bernstein inequality 24 and using the inequality $\sqrt{ab} \leqslant \frac{a}{H} + bH$:

$$\begin{aligned}
\big(p_h - \widehat{p}_h^t\big)\big(Z_{h+1}^\star - \mathring{Z}_{h+1}^t\big) &\leqslant \sqrt{2\,\mathrm{Var}_{\widehat{p}_h^t}(Z_{h+1}^\star - \mathring{Z}_{h+1}^t)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{2}{3}e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant \sqrt{2e^{\beta(H-h)}\widehat{p}_h^t(Z_{h+1}^\star - \mathring{Z}_{h+1}^t)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{2}{3}e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant \frac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t) + 2He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \frac{2}{3}e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}
\end{aligned}$$

Hence by combining the two bounds:

$$\widehat{p}_h^t\widetilde{Z}_{h+1}^t(s,a) - p_h\mathring{Z}_{h+1}^t(s,a) \leqslant 2b_h^t(s,a) + \Big(1 + \frac{2}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)$$

Hence by substituting and using lemma 11:

$$\widetilde{U}_h^t(s,a) - \mathring{U}_h^t(s,a) \leqslant e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1 + \frac{3}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)\Big]$$

Second case: $\mathring{U}_h^t(s,a) = \mathring{U}_{h,\mathrm{pes}}^t(s,a)$. In this case:

$$\begin{aligned}
\widetilde{U}_h^t(s,a) - \mathring{U}_h^t(s,a) &\leqslant e^{\beta r_h^t(s,a)}\Big[b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t)(s,a) + \widehat{p}_h^t\widetilde{Z}_{h+1}^t(s,a) \\
&\quad - \Big(\widehat{p}_h^t\mathring{Z}_{h+1}^t(s,a) - b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t)(s,a)\Big) \\
&= e^{\beta r_h^t(s,a)}\Big[2b_h^t(s,a) + \Big(1 + \frac{1}{H}\Big)\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t) + \frac{1}{H}(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t)\Big)(s,a)\Big]
\end{aligned}$$

Hence by lemma 11 we find:

$$\widetilde{U}_h^t(s,a) - \mathring{U}_h^t(s,a) \leqslant e^{\beta r_h^t(s,a)}\Big[2b_h^t(s,a) + \Big(1 + \frac{2}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)\Big]$$

Which conclude the recursion ■

We now prove lemma 10:

Proof We first prove by backward induction that, for all h and s,

$$\widetilde{Z}_h^t(s) - \mathring{Z}_h^t(s) \;\leqslant\; (\pi_h^{t+1}G_h^t)(s). \tag{15}$$

For $h = H+1$ it holds since both sides are 0. Assume it holds at step $h+1$. For $a = \pi_h^{t+1}(s)$

$$\widetilde{Z}_h^t(s) - \mathring{Z}_h^t(s) = \widetilde{U}_h^t(s,a) - \mathring{U}_h^t(s,a).$$

By lemma 12, we have :

$$\widetilde{U}_h^t(s,a) - \mathring{U}_h^t(s,a) \leqslant e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)\Big]$$

By the induction hypothesis inside the tilted expectation:

$$\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t)(s,a) \leqslant \widehat{p}_h^t(\pi^{t+1}G_{h+1}^t)(s,a).$$

Thus,

$$\widetilde{Z}_h^t(s) - \mathring{Z}_h^t(s) \leqslant 3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t(\pi^{t+1}G_{h+1}^t)(s,a) \leqslant G_h^t(s,a) = (\pi_h^{t+1}G_h^t)(s)$$

Finally, use optimism and the ring bridge: on $\mathcal{E}$, $Z_h^\star \leqslant \widetilde{Z}_h^t$ (lemma 9) and $\mathring{Z}_h^t \leqslant Z_h^{\pi^{t+1}}$ (lemma 11), hence

$$Z_h^\star(s) - Z_h^{\pi^{t+1}}(s) \leqslant \widetilde{Z}_h^t(s) - \mathring{Z}_h^t(s) \leqslant (\pi_h^{t+1}G_h^t)(s)$$

■

B.1.3. SAMPLE COMPLEXITY

Now we prove theorem 4 for $\beta > 0$

Proof the width certificate is:

$$G_h^t(s,a) = \min\left\{e^{\beta(H-h)}, e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t\pi^{t+1}G_{h+1}^t(s)\Big]\right\}$$

Let us transition to the true MDP. Using Bernstein inequality:

$$\big|(\widehat{p}_h^t - p_h)\pi^{t+1}G_{h+1}^t(s)\big| \leqslant \sqrt{2\operatorname{Var}_{p_h}\big(\pi^{t+1}G_{h+1}^t(s)\big)\alpha(n_h^t(s,a)} + \frac{2}{3}e^{\beta(H-h)}\alpha(n_h^t(s,a)$$

Now, we use the inequality $\operatorname{Var}(\pi_{h+1}^{t+1}G_{h+1}^t(s)) \leqslant e^{\beta(H-h)}\pi_{h+1}^{t+1}G_{h+1}^t(s)$. Hence, using the inequality $\sqrt{xy} \leqslant x+y$:

$$\big|(\widehat{p}_h^t - p_h)\pi^{t+1}G_{h+1}^t(s)\big| \leqslant \frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 3He^{\beta(H-h)}\alpha(n_h^t(s,a))$$

And using the variance transportation lemmas 27,28 and that $\alpha^\star(n_h^t(s,a)) \leqslant \alpha(n_h^t(s,a))$:

$$\begin{aligned}
\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\widetilde{Z}_{h+1}^t)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant 2\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \sqrt{4e^{\beta(H-h)}p_h(\widetilde{Z}_{h+1}^t - Z_{h+1}^{\pi^{t+1}})(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + 2e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant 2\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star)(s,a) + 2e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant 2\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}}r)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 2e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}
\end{aligned}$$

Hence, by coarsening the constants for the sake of simplicity:

$$\begin{aligned}
b_h^t(s,a) &\leqslant 4\sqrt{2}\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4(2\sqrt{2}+1)He^{\beta(H-h)}\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\
&\quad + (5+4\sqrt{2})e^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \frac{2\sqrt{2}}{H}p_h\pi^{t+1}G_{h+1}^t(s) \\
&\leqslant 6\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{2\sqrt{2}}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 27He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}
\end{aligned}$$

We combine the two terms:

$$\begin{aligned}
G_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\Bigg[&36\sqrt{\mathrm{Var}_{p_h}(Z_h^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{6\sqrt{2}}{H}p_h\pi^{t+1}G_{h+1}^t(s) \\
&+ 81He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \Big(1+\frac{3}{H}\Big)p_h\pi^{t+1}G_{h+1}^t(s) \\
&+ \Big(1+\frac{3}{H}\Big)\frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + \Big(1+\frac{3}{H}\Big)3He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\Bigg]
\end{aligned}$$

Hence, simplifying it gives:

$$G_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\left[36\sqrt{\mathrm{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \left(1+\frac{13}{H}\right)p_h\pi^{t+1}G_{h+1}^t(s) + 84He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\right]$$

Unrolling this inequality and using the terminal condition $G_{H+1}^t = 0$ we get:

$$(\pi_1 G_1^t)(s_1) \leqslant \mathbb{E}^{\pi}\Bigg[\sum_{h=1}^{H} \kappa^{h-1} \exp\Big(\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\Bigg(36\sqrt{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)\alpha^\star\big(n_h^t(s_h,a_h)\wedge 1\big)} \\ + 84He^{\beta(H-h)}\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)\Bigg)\Bigg| s_1\Bigg]$$

where $\kappa = 1 + \frac{13}{11}$. Since we have:

$$\kappa^{h-1} = \Big(1+\frac{13}{H}\Big)^{h-1} \leqslant \lim_{H\to+\infty}\Big(1+\frac{13}{H}\Big)^{H} = e^{13}$$

we get:

$$(\pi_1 G_1^t)(s_1) \leqslant e^{13}\mathbb{E}^{\pi^{t+1}}\Bigg[\sum_{h=1}^{H} \exp\Big(\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\Bigg(36\sqrt{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi}\right)\alpha^\star\big(n_h^t(s_h,a_h)\wedge 1\big)} \\ + 84He^{\beta(H-h)}\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)\Bigg)\Bigg| s_1\Bigg]$$

The algorithm stops when:

$$\pi_1^\tau G_1(s_1) \leqslant \frac{e^{|\beta|\varepsilon}-1}{e^{|\beta|\varepsilon}}\widetilde{Z}_1^\tau(s_1)$$

We upper bound $\frac{\pi_1^t G_1(s_1)}{\widetilde{Z}_1^t(s_1)}$ for $t = 1, ..., \tau-1$, using optimism:

$$\frac{\pi_1 G_1^t(s_1)}{\widetilde{Z}_1^t(s_1)} \leqslant \frac{(\pi_1 G_1^t(s_1)}{Z_1^{\pi^{t+1}}(s_1)} \\ \leqslant e^{13}\mathbb{E}^{\pi^{t+1}}\Bigg[\sum_{h=1}^{H} \exp\Big(\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\Bigg(36\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)(s_h,a_h)}{(Z_1^{\pi^{t+1}})^2(s_1)}\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)} \\ + 84He^{\beta(H-h)}\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)\Bigg)\Bigg| s_1\Bigg]$$

Let us bound the first term, using Cauchy-Schwartz inequality:

$$\mathbb{E}^{\pi^{t+1}}\Bigg[\sum_{h=1}^{H} \exp\Big(\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\Bigg(\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\big(\frac{\alpha^\star\big(n_h^t(s,a)}{n_h^t(s,a)}\wedge 1\big)}\Bigg)\Bigg| s_1\Bigg] \\ = \sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\Big(\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\Bigg(\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\big(\frac{\alpha^\star\big(n_h^t(s,a)}{n_h^t(s,a)}\wedge 1\big)}\Bigg) \\ \leqslant \sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\Big(2\beta \sum_{i=1}^{h} r_i(s_i,a_i)\Big)\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}}$$

For a policy π. By lemma 25 we have:

$$\sigma V_h^\pi(s) = e^{2\beta r_h(s,\pi(s))} \operatorname{Var}_{p_h}\left(Z_{h+1}^\pi\right)(s,\pi(s)) + e^{2\beta r_h(s,\pi(s))}\left(p_h \sigma V_{h+1}^\pi\right)(s,\pi(s))$$

Multiplying the equation by $\sum_{i=1}^{h-1} r_i(s_i,a_i)$ and take the expectation under π:

$$\mathbb{E}^\pi\left[e^{2\beta\sum_{i=1}^{h-1} r_i(s_i,a_i)}\sigma V_h^\pi(s_h)\right] = \mathbb{E}^\pi\left[e^{2\beta\sum_{i=1}^{h} r_i(s_i,a_i)}\operatorname{Var}_{p_h}\left(Z_{h+1}^\pi\right)\right] + \mathbb{E}^\pi\left[e^{2\beta\sum_{i=1}^{h} r_i(s_i,a_i)}\sigma V_{h+1}^\pi(s_{h+1})\right]$$

By summing over $h = 1, ..., H$ and since $\sigma V_{H+1}^\pi = 0$ we get:

$$\sum_{h=1}^{H}\mathbb{E}^\pi\left[e^{2\beta\sum_{i=1}^{h} r_i(s_i,a_i)}\frac{\operatorname{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\right] = \mathbb{E}^\pi\left[\frac{\sigma V_1^\pi(s_1)}{(Z_1^\pi)^2}\right]$$

But notice that for a deterministic policy π:

$$\frac{\sigma V_1^\pi(s_1)}{(Z_1^\pi)^2} = \frac{\operatorname{Var}(e^{\beta R_1^\pi}|S_1 = s_1)}{\mathbb{E}(e^{\beta R_1^\pi}|S_1 = s_1)^2}$$

Using lemma 26 we get that:

$$\frac{\sigma V_1^\pi(s_1)}{(Z_1^\pi)^2} \leqslant \frac{(e^{\beta G_{\max}(\mathcal{M})} - 1)^2}{e^{\beta G_{\max}(\mathcal{M})}}$$

Applying this for π^{t+1} we get:

$$\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\left(2\beta\sum_{i=1}^{h} r_i(s_i,a_i)\right)\frac{\operatorname{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2} \leqslant \frac{(e^{\beta G_{\max}(\mathcal{M})} - 1)^2}{e^{\beta G_{\max}(\mathcal{M})}}$$

For the second term:

$$\begin{aligned}
&\mathbb{E}^{\pi^{t+1}}\left[\sum_{h=1}^{H}\exp\left(\beta\sum_{i=1}^{h} r_i(s_i,a_i)\right)He^{\beta(H-h)}\left(\frac{\alpha\left(n_h^t(s,a)\right)}{n_h^t(s,a)} \wedge 1\right)\middle| s_1\right] \\
&= \sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\left(\beta\sum_{i=1}^{h} r_i(s_i,a_i)\right)He^{\beta(H-h)}\left(\frac{\alpha\left(n_h^t(s,a)\right)}{n_h^t(s,a)} \wedge 1\right) \\
&\leqslant \sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)e^{\beta h}He^{\beta(H-h)}\left(\frac{\alpha\left(n_h^t(s,a)\right)}{n_h^t(s,a)} \wedge 1\right) \\
&\leqslant He^{\beta H}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\left(\frac{\alpha\left(n_h^t(s,a)\right)}{n_h^t(s,a)} \wedge 1\right)
\end{aligned}$$

Hence:

$$
\begin{aligned}
(\pi_1 G_1^t)(s_1) &\leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}e^{\beta H}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big) \\
&\leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}}\sqrt{\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}e^{\beta H}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)
\end{aligned}
$$

Let us sum over $t \leqslant \tau$. By sub-optimality for each episode $t = 0, ..., \tau - 1$ we have:

$$
\pi_1^{t+1}G_1(s_1) > \frac{e^{|\beta|\varepsilon}-1}{e^{|\beta|\varepsilon}}\widetilde{Z}_1^t(s_1)
$$

Hence by summing over all the episodes and using Cauchy-Schwartz:

$$
\begin{aligned}
\tau\frac{e^{\beta\varepsilon}-1}{e^{\beta\varepsilon}} &\leqslant 36e^{13}\sum_{t=1}^{\tau-1}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}e^{\beta H}\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big) \\
&\leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}}\sqrt{T}\sqrt{\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}e^{\beta H}\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)
\end{aligned}
$$

Using lemma 29 to relate the true counts to pseudo-counts we get:

$$
\begin{aligned}
\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big) &\leqslant \sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\alpha\big(\widetilde{n}_h^t(s,a)\vee 1\big) \\
&\leqslant \alpha(\tau-1,\delta)\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\frac{1}{\widetilde{n}_h^t(s,a)\vee 1} \\
&\leqslant \alpha(\tau-1,\delta)\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}\frac{\widetilde{n}_{h+1}^t(s,a)-\widetilde{n}_h^t(s,a)}{\widetilde{n}_h^t(s,a)\vee 1} \\
&\leqslant 3SAH\alpha(\tau-1,\delta)\log(\tau+1)
\end{aligned}
$$

Where in the final inequality we used lemma 31. Similarly, we find:

$$\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\Big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)} \wedge 1\Big) \leqslant 3SAH\alpha^\star(\tau-1,\delta)\log(\tau+1)$$

Hence:

$$\tau\frac{e^{\beta\varepsilon}-1}{e^{\beta\varepsilon}} \leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\tau SAH\alpha^\star(t-1,\delta)\log(t+1)} + 84e^{13}e^{\beta H}SAH\alpha(t-1,\delta)\log(t+1)$$

Therefore, by replacing α^* and α by their expression and using that $\log(\tau+1) \leqslant \log(8e\tau)$ since $\tau \geqslant 1$:

$$\begin{aligned}\tau\frac{e^{\beta\varepsilon}-1}{e^{\beta\varepsilon}} \leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\tau SAH\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8e\tau\big)+\log\big(8e\tau\big)^2\Big)}\\ + 84e^{13}e^{\beta H}SAH\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8e\tau\big)+S\log\big(8e\tau\big)^2\Big)\end{aligned}$$

Finally, we use lemma 31 with :

$$C = 36e^{13}\frac{e^{\beta\varepsilon}}{e^{\beta\varepsilon}-1}\sqrt{\big(\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}SAH} \quad , A = \log\big(\frac{3SAH}{\delta}\big) \quad , B = 1$$
$$D = \frac{e^{\beta\varepsilon}}{e^{\beta\varepsilon}-1}84e^{13}e^{\beta H}H^2SA \quad \text{and} \quad E = S$$

Which yield:

$$\tau \leqslant \frac{e^{2\beta\varepsilon}}{\big(e^{\beta\varepsilon}-1\big)^2}\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}SAH\Big(\log\big(\frac{3SAH}{\delta}\big)+1\Big)C_1^2 + 3e^{\beta\varepsilon}\frac{e^{\beta H}H^2SA}{e^{\beta\varepsilon}-1}\Big(\log\big(\frac{3SAH}{\delta}\big)+S\Big)C_1^2+1$$

Where $C_1 = \frac{8}{5}\log\left(4e^{17}\frac{(S+1)(H+1)e^{|\beta H}SAH^2}{e^{\beta\varepsilon}-1}\right)$

In particular, if ε is small enough so that the first term dominates the second then:

$$\tau \leqslant \frac{e^{2\beta\varepsilon}}{\big(e^{\beta\varepsilon}-1\big)^2}\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}SAH\log\big(\frac{3SAH}{\delta}\big)C_2^2$$

Where $C_2 = 3eC_1$. We can finally hide the constants and the log terms to get:

$$\tau = \tilde{\mathcal{O}}\left(\frac{1}{\big(e^{\beta\varepsilon}-1\big)^2}\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}SAH\right)$$

Finally to see that the stopping rule implies (ε,δ)PAC, remark that at time τ:

$$\pi_1^\tau G_1(s_1) \leqslant \frac{e^{\beta\varepsilon}-1}{e^{\beta\varepsilon}}\widetilde{Z}_1^t(s_1)$$

This is equivalent to :

$$\pi_1^\tau G_1(s_1) \leqslant (e^{|\beta|\varepsilon} - 1)\left(\widetilde{Z}_1^t(s_1) - \pi_1^\tau G_1(s_1)\right)$$

Since $\widetilde{Z}_1^t(s_1) - Z_1^\star(s_1) \leqslant \widetilde{Z}_1^t(s_1) - \mathring{Z}_1^t(s_1) \leqslant \pi_{t+1,1} G_1^t(s_1)$, this stopping condition is stronger than the condition:

$$\pi_1^\tau G_1(s_1) \leqslant (e^{\beta\varepsilon} - 1) Z_1^{\pi^\tau}$$

But we can write:

$$(V_1^\star - V_1^{\pi^{\tau+1}})(s_1) = \frac{1}{\beta}\log\left(\frac{Z_1^\star}{Z_1^{\pi^\tau}}\right)(s_1) = \frac{1}{\beta}\log\left(1 + \frac{Z_1^\star - Z_1^{\pi^\tau}}{Z_1^{\pi^\tau}}\right)(s_1) \leqslant \frac{1}{\beta}\log\left(1 + \frac{\pi_1^\tau G_1(s_1)}{Z_1^{\pi^\tau}}\right)(s_1) \leqslant \varepsilon$$

■

B.2. Case $\beta < 0$

B.2.1. STOPPING RULE

We first discuss the stopping rule for $\beta < 0$. The difference in value functions can be expressed in the exponential space as:

$$\begin{aligned}(V_1^\star - V_1^{\pi^{t+1}})(s_1) &= \frac{1}{\beta}\log\left(\frac{Z_1^\star}{Z_1^{\pi^{t+1}}}\right)(s_1) \\ &= \frac{1}{\beta}\log\left(1 + \frac{Z_1^\star - Z_1^{\pi^{t+1}}}{Z_1^{\pi^{t+1}}}\right)(s_1).\end{aligned}$$

To ensure the policy is ε-optimal, it suffices to satisfy at stopping time τ:

$$\pi_1^t G_1(s_1) \leqslant \left(e^{\beta\varepsilon} - 1\right) Z_1^{\pi^t}(s_1) \tag{16}$$

However, $Z_1^{\pi^t}$ is unknown as it relies on the true dynamics. We therefore substitute it with a computable lower bound:

$$\pi_1^t G_1(s_1) \leqslant \left(e^{\beta\varepsilon} - 1\right) \underset{\sim}{Z}_1^{\pi^t}(s_1) \tag{17}$$

where both $\pi_1^t G_1(s_1)$ and $\underset{\sim}{Z}_1^t(s_1)$ can be computed efficiently by dynamic programming.

B.2.2. CONFIDENCE BOUNDS

We first start with a lemma:

Lemma 13 *On the good event $\mathcal{E}^-$:*

$$\begin{aligned}|\big(p_h - \widehat{p}_h^t\big) Z_{h+1}^\star| \leqslant 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_h^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5(1 - e^{\beta(H-h)})\alpha(n_h^t(s,a)) \\ + 4H(1 - e^{\beta(H-hH)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t\Big(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\end{aligned}$$

Proof On the good event $\mathcal{E}^-$ we have:

$$\left|(p_h - \widehat{p}_h^t)Z_{h+1}^\star(s,a)\right| \leqslant \sqrt{2\operatorname{Var}_{p_h}(Z_{h+1}^*)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 3(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}$$

Since $Z_h^\star$ and p_h are unknown, we use a variance transportation inequality to replace with its value for $\widetilde{Z}_h^t$ the optimistic bound for $V^\star$. By applying lemma 27 and lemma 28 successively:

$$\begin{aligned}\operatorname{Var}_{p_h}(Z_{h+1}^\star) &\leqslant 2\operatorname{Var}_{\widehat{p}_h^t}(Z_{h+1}^\star)(s,a) + 4(1-e^{\beta(H-h)})^2\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\ &\leqslant 4\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_{h+1}^t)(s,a) + 4(1-e^{\beta(H-h)})\widehat{p}_h^t(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t)(s,a) + 4(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\end{aligned}$$

Hence, using the inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ and then $\sqrt{ab} \leqslant a\eta + \frac{b}{\eta}$ for $\eta = H$ and using that $\alpha^\star(n_h^t(s,a)) \leqslant \alpha(n_h^t(s,a))$:

$$\begin{aligned}\sqrt{\operatorname{Var}_{p_h}(Z_h^\star)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant 2\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_h^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\ &\quad + \sqrt{4(1-e^{\beta(H-h)})\widehat{p}_h^t(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\ &\quad + 2(1-e^{\beta(H-h)})\sqrt{\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} \\ &\leqslant 2\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_h^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4H(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\ &\quad + \frac{1}{H}\widehat{p}_h^t(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t)(s,a) + 2(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\end{aligned}$$

Hence:

$$\begin{aligned}|p_h Z_{h+1}^\star - \widehat{p}_h^t(s,a)Z_{h+1}^\star| &\leqslant 2\sqrt{2}\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_{h+1}^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5(1-e^{\beta(H-h)})\alpha(n_h^t(s,a)) \\ &\quad + 4H(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_h^t\left(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t\right)(s,a)\end{aligned}$$

■

Denote the bonus term as:

$$b_h^t(s,a) = 2\sqrt{2}\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_h^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 5(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + 4H(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \tag{18}$$

Now, like the case $\beta > 0$ we define define optimistic and pessimistic state-value function on the exponential transform of $Q_h^\star$ which denoted by $U_h^\star$.

$$\begin{aligned}
\widetilde{U}_h^t(s,a) &= \min\left\{1, e^{\beta r_h(s,a)}\left[\widehat{p}_h^t \widetilde{Z}_{h+1}^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]\right\} \\
\widetilde{Z}_h^t(s) &= \min_{a\in\mathcal{A}} \widetilde{U}_h^t(s,a), \qquad \widetilde{Z}_{H+1}^t(s) = 1 \\
\underset{\sim}{U}_h^t(s,a) &= \max\left\{e^{\beta(H-h-1)}, e^{\beta r_h(s,a)}\left[\widehat{p}_h^t \underset{\sim}{Z}_{h+1}^t(s,a) - b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]\right\} \\
\underset{\sim}{Z}_h^t(s) &= \min_{a\in\mathcal{A}} \underset{\sim}{U}_h^t(s,a), \qquad \underset{\sim}{Z}_{H+1}^t(s) = 1
\end{aligned} \tag{19}$$

And we consider the greedy policy:

$$\pi_h^{t+1}(s) = \arg\min_{a\in\mathcal{A}} \underset{\sim}{U}_h^t(s,a)$$

Let us prove an optimism lemma:

Lemma 14 *On the good event $\mathcal{E}^-$ we have:*

$$\underset{\sim}{U}_h^t(s,a) \leqslant U_h^\star(s,a) \leqslant \widetilde{U}_h^t(s,a)$$

and

$$\underset{\sim}{Z}_h^t(s) \leqslant Z_h^\star(s) \leqslant \widetilde{Z}_h^t(s)$$

Proof We proceed by induction over h. For $h = H+1$ the result is trivially upper bounding and (resp. lower bounding) $Q^\star$ by H and 1.
Assume the inequality holds for $h' > h$. Fix (s,a) and assume $\widetilde{Q}_h^t(s,a) < H$ .we have that:

$$\begin{aligned}
\widetilde{U}_h^t(s,a) - U_h^\star(s,a) &= e^{\beta r_h(s,a)}\left[\widehat{p}_h^t \widetilde{Z}_{h+1}^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) - p_h Z_{h+1}^\star(s,a)\right] \\
&= e^{\beta r_h(s,a)}\Bigg[\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t(s,a) - Z_{h+1}^\star(s,a)\Big) + \Big(\widehat{p}_h^t - p_h\Big)Z_{h+1}^\star(s,a) \\
&\qquad + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\Bigg]
\end{aligned}$$

But we know by Bernstein inequality that:

$$\Big(\widehat{p}_h^t - p_h\Big)Z_h^\star(s,a) \geqslant -b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(Z_{h+1}^\star - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) \geqslant -b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a)$$

Hence:

$$\widetilde{U}_h^t(s,a) - U_h^\star(s,a) \geqslant e^{\beta r_h(s,a)}\left[\Big(1 + \frac{1}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t(s,a) - Z_{h+1}^\star(s,a)\Big)\right] \geqslant 0$$

Where we used the induction hypothesis. We prove the pessimistic property in the same way ■

B.2.3. STOPPING RULE

We define the stopping rule for the algorithm for $\beta < 0$ as:

$$G_h^t(s,a) = \min\left\{1, e^{\beta r_h(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t \pi^{t+1} G_{h+1}^t(s,a)\Big]\right\} \tag{20}$$

Lemma 15 establishes the validity of this stopping rule by showing that, with high probability, it bounds the certificate width:

Lemma 15 *On the good event $\mathcal{E}^-$, for all t and all h,*

$$Z_h^{\pi^{t+1}}(s) - Z_h^\star(s) \leqslant \pi_h^{t+1} G_h^t(s) \qquad \forall s \in \mathcal{S}$$

In particular, at the initial state s_1, $V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1) \leqslant \pi_1^{t+1} G_1^t(s_1)$

We prove the lemma 15 in this section : As in the case $\beta < 0$. We define the auxiliary (analysis-only) variable $\mathring{Z}_h^t$. Setting $\mathring{Z}_{H+1}^t \equiv 1$, we recurse backward for $h = H, \dots, 1$:

$$\begin{aligned}
\mathring{U}_{h,\text{opt}}^t(s,a) &= \min\Big\{1, e^{\beta r_h(s,a)}\Big[(\widehat{p}_h^t \mathring{Z}_{h+1}^t)(s,a) + b_h^t(s,a) + \tfrac{1}{H}\big(\widehat{p}_h^t(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t)\big)(s,a)\Big]\Big\},\\
\mathring{U}_h^t(s,a) &= \max\Big\{e^{\beta r_h(s,a)}(p_h \mathring{Z}_{h+1}^t)(s,a), \mathring{U}_{h,\text{opt}}^t(s,a)\Big\},\\
\mathring{Z}_h^t(s) &= \mathring{U}_h^t\big(s, \pi_h^{t+1}(s)\big).
\end{aligned}$$

Because $\widetilde{Z}^t$ is pessimistic with respect to $Z^\star$ (and $\beta < 0$ reverses the relevant order), we cannot directly compare it to $Z^{\pi^{t+1}}$. We introduce $\mathring{Z}^t$ as a bridge quantity. Intuitively, $\mathring{Z}^t$ satisfies the exponential Bellman recursion under the true kernel p_h while being clipped by an optimistic empirical backup; hence it serves as a worst-case upper bound for both $\widetilde{Z}^t$ and $Z^{\pi^{t+1}}$.

Lemma 16 *For all t, For all (h,s,a):*

$$\mathring{U}_h^t(s,a) \geqslant \max\Big(\widetilde{U}_h^t(s,a), U_h^{\pi_h^{t+1}}(s,a)\Big)$$

and

$$\mathring{Z}_h^t(s) \geqslant \max\Big(\widetilde{Z}_h^t(s), Z_h^{\pi_h^{t+1}}(s,a)\Big)$$

Proof We proceed by backward induction. For $h = H+1$, all values are equal to 0 so the inequalities hold. Assume that for some $h \leqslant H$ we have for all (s,a):

$$\mathring{U}_{h+1}^t(s,a) \geqslant \max\Big(\widetilde{U}_{h+1}^t(s,a), U_{h+1}^{\pi^{t+1}h}(s,a)\Big)$$

and

$$\mathring{Z}_{h+1}^t(s) \geqslant \max\Big(\widetilde{Z}_{h+1}^t(s), Z_{h+1}^{\pi^{t+1}}(s,a)\Big)$$

we have by construction:

$$\mathring{U}_h^t(s,a) \geqslant \mathring{U}_{h,\text{true}}^t(s,a) = e^{\beta r_h(s,a)}(p_h \mathring{Z}_{h+1}^t)(s,a) \geqslant e^{\beta r_h(s,a)}(p_h Z_{h+1}^{\pi^{t+1}})(s,a) = U_h^{\pi^{t+1}}(s,a)$$

Where we used the induction hypothesis and the monotonicity of the exponential Bellman operator.

$$\begin{aligned}
\mathring{U}_h^t(s,a) - \widetilde{U}_h^t(s,a) &\geqslant \mathring{U}_{h,\text{opt}}^t(s,a) - \widetilde{U}_h^t(s,a) \\
&\geqslant e^{\beta r_h(s,a)}\Bigg[\widehat{p}_h^t \mathring{Z}_{h+1}^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) \\
&\quad - \Big(\widehat{p}_h^t \widetilde{Z}_h^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\Big)\Bigg] \\
&\geqslant e^{\beta r_h(s,a)}\left[\left(1+\frac{1}{H}\right)\widehat{p}_h^t\left(\mathring{Z}_{h+1}^t - \widetilde{Z}_{h+1}^t\right)(s,a)\right]
\end{aligned}$$

Where we applied the induction hypothesis, we conclude then:

$$\mathring{U}_h^t(s,a) \geqslant \widetilde{U}_h^t(s,a)$$

For Z:

$$\mathring{Z}_h^t(s) - Z_h^{\pi^{t+1}}(s,a) = \mathring{U}_h^t(s,\pi_h^t(s)) - U_h^{\pi^{t+1}}(s,\pi_h^{t+1}(s)) \geqslant 0$$

And:

$$\mathring{Z}_h^t(s) = \mathring{U}_h^t(s,\pi_h^t(s)) \geqslant \widetilde{U}_h^t(s,\pi_h^t(s)) \geqslant \min_{a\in\mathcal{A}} \widetilde{U}_h^t(s,a) = \widetilde{Z}_h^t(s,a)$$

Which conclude the recurrence ■

Lemma 17 *For any t, any $h \in \{1,...,H\}$ and any state-action pair:*

$$\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t\Big(\mathring{Z}_h^t(s,a) - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\Big]$$

Proof Fix $h \in \{1,...,H\}$ and fix a state-action pair (s,a). We have two cases:
First case: $\mathring{U}_h^t(s,a) = \mathring{U}_{h,\text{true}}^t(s,a)$, we have:

$$\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\Big[b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) + p_h\mathring{Z}_{h+1}^t(s,a) - \widehat{p}_h^t\underset{\sim}{Z}_{h+1}^t(s,a)\Big]$$

The last term can be written as:

$$p_h\mathring{Z}_{h+1}^t(s,a) - \widehat{p}_h^t\underset{\sim}{Z}_{h+1}^t(s,a) = \widehat{p}_h^t\Big(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) + \big(\widehat{p}_h^t - p_h\big)Z_{h+1}^\star(s,a) + \big(p_h - \widehat{p}_h^t\big)\big(\mathring{Z}_{h+1}^t - Z_{h+1}^\star\big)$$

For the second term, by lemma 13:

$$\big|\big(p_h - \widehat{p}_h^t\big)Z_{h+1}^\star(s,a)\big| \leqslant b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - Z_{h+1}^\star\Big)(s,a) \leqslant b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)(s,a)$$

For the third term, by the KL-Bernstein inequality 24 and using the inequality $\sqrt{ab} \leqslant \frac{a}{H} + bH$:

$$\begin{aligned}
\big(p_h - \widehat{p}_h^t\big)\big(Z_{h+1}^\star - \mathring{Z}_{h+1}^t\big) &\leqslant \sqrt{2\,\mathrm{Var}_{\widehat{p}_h^t}(Z_{h+1}^\star - \mathring{Z}_{h+1}^t)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{2}{3}\left(1 - e^{\beta(H-h)}\right)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant \sqrt{2e^{\beta(H-h)}\widehat{p}_h^t(Z_{h+1}^\star - \mathring{Z}_{h+1}^t)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{2}{3}\left(1 - e^{\beta(H-h)}\right)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant \frac{1}{H}\widehat{p}_h^t(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t) + 2He^{\beta(H-h)}\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \frac{2}{3}\left(1 - e^{\beta(H-h)}\right)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}
\end{aligned}$$

Hence by combining the two bounds:

$$p_h\mathring{Z}_{h+1}^t(s,a) - \widehat{p}_h^t\underset{\sim}{Z}_{h+1}^t(s,a) \leqslant 2b_h^t(s,a) + \Big(1 + \frac{2}{H}\Big)\widehat{p}_h^t\Big(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)$$

Hence by substituting and using lemma 16:

$$\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) \leqslant e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1 + \frac{3}{H}\Big)\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \mathring{Z}_{h+1}^t\Big)\Big]$$

Second case: $\mathring{U}_h^t(s,a) = \mathring{U}_{h,\mathrm{opt}}^t(s,a)$, we have:

$$\begin{aligned}
\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) &\leqslant e^{\beta r_h(s,a)}\Bigg[\widehat{p}_h^t\mathring{Z}_{h+1}^t(s,a) + b_h^t(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) \\
&\quad - \left(\widehat{p}_h^t\underset{\sim}{Z}_{h+1}^t(s,a) - b_h^t(s,a) - \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right)\Bigg] \\
&= e^{\beta r_h(s,a)}\left[2b_h^t(s,a) + \left(1 + \frac{1}{H}\right)\widehat{p}_h^t\Big(\mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) + \frac{1}{H}\widehat{p}_h^t\Big(\widetilde{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\right]
\end{aligned}$$

Using lemma 16 we get:

$$\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\Big[2b_h^t(s,a) + \Big(1 + \frac{2}{H}\Big)\widehat{p}_h^t\Big(\mathring{Z}_h^t(s,a) - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\Big]$$

■

We now prove lemma 15:

Proof We first prove by backward induction that, for all h and s,

$$\mathring{Z}_h^t(s) - \underset{\sim}{Z}_h^t(s) \;\leqslant\; (\pi_h^{t+1}G_h^t)(s).$$

For $h = H+1$ it holds since both sides are 0. Assume it holds at step $h+1$. For $a = \pi_h^{t+1}(s)$

$$\mathring{Z}_h^t(s) - \underset{\sim}{Z}_h^t(s) = \mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a)$$

Apply Lemma 16:

$$\mathring{U}_h^t(s,a) - \underset{\sim}{U}_h^t(s,a) \leqslant e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1 + \frac{3}{H}\Big)\widehat{p}_h^t\Big(\mathring{Z}_h^t(s,a) - \underset{\sim}{Z}_{h+1}^t\Big)(s,a)\Big]$$

By the induction hypothesis:

$$\widehat{p}_h^t\Big(\mathring{Z}_h^t(s,a) - \underset{\sim}{Z}_{h+1}^t\Big)(s,a) \leqslant \widehat{p}_h^t(\pi^{t+1}G_{h+1}^t)(s,a).$$

Thus,

$$\mathring{Z}_h^t(s) - \underset{\sim}{Z}_h^t(s) \leqslant 3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t(\pi^{t+1}G_{h+1}^t)(s,a) \leqslant G_h^t(s,a) = (\pi_h^{t+1}G_h^t)(s)$$

Finally, use optimism and the ring bridge: on $\mathcal{E}$, $Z_h^\star \geqslant \underset{\sim}{Z}_h^t$ (optimism lemma 14) and $Z_h^{\pi^{t+1}} \leqslant \mathring{Z}_h^t$ (Lemma 16):

$$Z_h^{\pi^{t+1}}(s) - Z_h^\star(s) \leqslant \mathring{Z}_h^t(s) - \underset{\sim}{Z}_h^t(s) \leqslant (\pi_h^{t+1}G_h^t)(s)$$

■

B.2.4. SAMPLE COMPLEXITY

Proof the width certificate is:

$$G_h^t(s,a) = \min\left\{1, e^{\beta r_h^t(s,a)}\Big[3b_h^t(s,a) + \Big(1+\frac{3}{H}\Big)\widehat{p}_h^t\pi^{t+1}G_{h+1}^t(s)\Big]\right\}$$

Let us transition to the true MDP. Using Bernstein inequality:

$$\big|(\widehat{p}_h^t - p_h)\pi^{t+1}G_{h+1}^t(s)\big| \leqslant \sqrt{2\operatorname{Var}_{p_h}\left(\pi^{t+1}G_{h+1}^t(s)\right)\frac{\alpha(n_h^t(s,a)}{n_h^t(s,a)}} + \frac{2}{3}(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}$$

Now, we use the inequality $\operatorname{Var}(\pi_{h+1}^{t+1}G_{h+1}^t(s)) \leqslant (1-e^{\beta(H-h)})\pi_{h+1}^{t+1}G_{h+1}^t(s)$. Hence, using the inequality $\sqrt{xy} \leqslant x+y$:

$$\big|(\widehat{p}_h^t - p_h)\pi^{t+1}G_{h+1}^t(s)\big| \leqslant \frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 3H(1-e^{\beta(H-h+1)})\frac{\alpha(n_h^t(s,a)}{n_h^t(s,a)}$$

And using the variance transportation lemmas 27 and 28 and that $\alpha^\star(n_h^t(s,a)) \leqslant \alpha(n_h^t(s,a))$:

$$\begin{aligned}
\sqrt{\operatorname{Var}_{\widehat{p}_h^t}(\underset{\sim}{Z}_{h+1}^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} &\leqslant 2\sqrt{\operatorname{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + \sqrt{4(1-e^{\beta(H-h)})p_h(Z_{h+1}^{\pi^{t+1}} - \underset{\sim}{Z}_{h+1}^t)(s,a)\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} \\
&\quad + 2(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant 2\sqrt{\operatorname{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4H(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h(Z_{h+1}^{\pi^{t+1}} - \underset{\sim}{Z}_{h+1}^t)(s,a) + 2(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} \\
&\leqslant 2\sqrt{\operatorname{Var}_{p_h}(Z_{h+1}^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + 4H(1-e^{\beta(H-h)})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)} \\
&\quad + \frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 2(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}
\end{aligned}$$

Since we have :

$$Z_{h+1}^{\pi^{t+1}} - \underset{\sim}{Z}_{h+1}^t \leqslant \mathring{Z}_{h+1}^t - \underset{\sim}{Z}_{h+1}^t \leqslant G_{h+1}^t(s)$$

Hence, using that $\alpha^\star(n_h^t(s,a)) \leqslant \alpha(n_h^t(s,a))$ and simplifying constants and using that $H \geqslant 1$:

$$b_h^t(s,a) \leqslant 6\sqrt{\mathrm{Var}_{p_h}(Z_h^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{3}{H}p_h\pi^{t+1}G_{h+1}^t(s) + 27H(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}$$

We combine the two terms:

$$\begin{aligned} G_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\Bigg[& 36\sqrt{\mathrm{Var}_{p_h}(Z_h^{\pi^{t+1}})\frac{\alpha^\star(n_h^t(s,a))}{n_h^t(s,a)}} + \frac{6\sqrt{2}}{H}p_h\pi^{t+1}G_{h+1}^t(s) \\ & + 81H(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)} + \Big(1+\frac{3}{H}\Big)p_h\pi^{t+1}G_{h+1}^t(s) \\ & + \Big(1+\frac{3}{H}\Big)\frac{1}{H}p_h\pi^{t+1}G_{h+1}^t(s) + \Big(1+\frac{3}{H}\Big)3H(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\Bigg] \end{aligned}$$

Hence, simplifying it gives:

$$G_h^t(s,a) \leqslant e^{\beta r_h(s,a)}\left[36\sqrt{\mathrm{Var}_{p_h}(Z_h^{\pi^{t+1}})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}} + \left(1+\frac{13}{H}\right)p_h\pi^{t+1}G_{h+1}^t(s) + 81H(1-e^{\beta(H-h)})\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}\right]$$

Unrolling this inequality like the case $\beta > 0$:

$$\begin{aligned} (\pi_1 G_1^t)(s_1) \leqslant e^{13}\mathbb{E}^\pi\Bigg[& \sum_{h=1}^H \exp\Big(\beta\sum_{i=1}^h r_i(s_i,a_i)\Big)\Bigg(36\sqrt{\mathrm{Var}_{p_h}\left(Z_{h+1}^\pi\right)\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)} \\ & + 81H(1-e^{\beta(H-h)})\alpha\big(n_h^t(s_h,a_h)\wedge 1\big)\Bigg)\Bigg| s_1\Bigg] \end{aligned}$$

The algorithm stops when:

$$\pi_1^\tau G_1(s_1) \leqslant (1-e^{\beta\varepsilon})\underset{\sim}{Z}_1^{\pi^\tau}$$

This is equivalent to:

$$\pi_1^\tau G_1(s_1) \leqslant \frac{e^{|\beta|\varepsilon}-1}{2e^{|\beta|\varepsilon}-1}\left(\underset{\sim}{Z}_1^{\pi^\tau} + \pi_1 G_1^t(s_1)\right)$$

We then need to upper bound the quantity $\frac{\pi_1^t G_1(s_1)}{\underset{\sim}{Z}_1^{\pi^t}}$ for $t = 1, ..., \tau-1$:

$$\begin{aligned} \frac{\pi_1 G_1^t(s_1)}{\underset{\sim}{Z}_1^{\pi^t} + \pi_1 G_1^t(s_1)} & \leqslant \frac{\pi_1 G_1^t(s_1)}{Z_1^{\pi^{t+1}}(s_1)} \\ & \leqslant \frac{e^{13}}{\beta}\mathbb{E}^{\pi^{t+1}}\Bigg[\sum_{h=1}^H \exp\Big(\beta\sum_{i=1}^h r_i(s_i,a_i)\Big)\Bigg(36\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\Big(\frac{\alpha\big(n_h^t(s_h,a_h)\big)}{n_h^t(s,a)}\wedge 1\Big)} \\ & \quad + 81H(1-e^{\beta(H-h)})\Big(\frac{\alpha\big(n_h^t(s_h,a_h)\big)}{n_h^t(s,a)}\wedge 1\Big)\Bigg)\Bigg| s_1\Bigg] \end{aligned}$$

Like the case $\beta > 0$ we write directly:

$$\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\Big(2\beta\sum_{i=1}^{h} r_i(s_i,a_i)\Big)\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi}\right)}{(Z_1^{\pi^{t+1}})^2} = \mathbb{E}^{\pi}\big[\frac{\sigma V_1^{\pi^{t+1}}}{(Z_1^{\pi^{t+1}})^2}\big]$$

But since the greedy policy is deterministic we have:

$$\frac{\sigma V_1^{\pi}(s_1)}{(Z_1^{\pi})^2} = \frac{\mathrm{Var}(e^{\beta R_1^{\pi}}|S_1 = s_1)}{\mathbb{E}(e^{\beta R_1^{\pi}}|S_1 = s_1)^2}$$

Where $L = e^{\beta\sum_{i=1}^{H} r_i(s_i,a_i)}$, using lemma 26 we get that:

$$\frac{\sigma V_1^{\pi}(s_1)}{(Z_1^{\pi})^2} \leqslant \frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}$$

Hence:

$$\begin{aligned}
&\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\Big(\beta\sum_{i=1}^{h} r_i(s_i,a_i)\Big)\left(36\sqrt{\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)}\right)\\
&\quad\leqslant 36\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\Big(2\beta\sum_{i=1}^{h} r_i(s_i,a_i)\Big)\frac{\mathrm{Var}_{p_h}\left(Z_{h+1}^{\pi^{t+1}}\right)}{(Z_1^{\pi^{t+1}})^2}}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\alpha(n_h^t(s,a))}{n_h^t(s,a)}}\\
&\quad\leqslant 36\sqrt{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}}\sqrt{\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\frac{\alpha^{\star}(n_h^t(s,a))}{n_h^t(s,a)}}
\end{aligned}$$

For the second term:

$$\begin{aligned}
&\frac{1}{Z_1^{\pi^{t+1}}(s_1)}\mathbb{E}^{\pi^{t+1}}\left[\sum_{h=1}^{H}\exp\left(\beta\sum_{i=1}^{h} r_i(s_i,a_i)\right)H(1-e^{\beta(H-h)})\left(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\right)\middle| s_1\right]\\
&\quad= \frac{1}{Z_1^{\pi^{t+1}}}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\exp\left(\beta\sum_{i=1}^{h} r_i(s_i,a_i)\right)H(1-e^{\beta(H-h)})\left(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\right)\\
&\quad\leqslant \sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)e^{\beta h}H(1-e^{\beta(H-h)})\left(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\right)\\
&\quad\leqslant He^{|\beta|H}\sum_{h=1}^{H}\sum_{s,a} p_h^{t+1}(s,a)\left(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\right)
\end{aligned}$$

We then sum on $t < \tau$ and use that by sub-optimality we have for $t = 1, ..., \tau - 1$:

$$\frac{\pi_1 G_1^t(s_1)}{Z_1^{\pi^{t+1}}} \geqslant \frac{\pi_1^{\tau} G_1(s_1)}{(Z_1^{\pi^{\tau}} + \pi_1 G_1^t(s_1))} \geqslant e^{|\beta|\varepsilon} - 1$$

Hence:

$$\tau(e^{|\beta|\varepsilon}-1) \leqslant 36e^{13}e^{2|\beta|\varepsilon}\sum_{t=1}^{\tau-1}\sqrt{(\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}})\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1)}$$
$$+ 81e^{13}e^{2|\beta|\varepsilon}He^{|\beta|H}\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1)$$
$$\leqslant 36e^{13}e^{2|\beta|\varepsilon}\sqrt{(\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}})}\sqrt{T}\sqrt{\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1)}$$
$$+ 81e^{13}e^{2|\beta|\varepsilon}He^{|\beta|H}\sum_{t=1}^{\tau-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1)$$

Similarly to the case $\beta > 0$ we bound the other terms using the counting argument which yield:

$$\tau(e^{|\beta|\varepsilon}-1) \leqslant 36e^{13}e^{2|\beta|\varepsilon}\sqrt{(\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}})\tau SAH\alpha^\star(\tau-1,\delta)\log(\tau+1)} + 81e^{13}e^{2|\beta|\varepsilon}e^{|\beta|H}H^2SA\alpha(\tau-1,\delta)$$

We replace $\alpha^\star$ and α by their expressions and using that $\log(\tau+1)\leqslant\log(8e\tau)$ since $\tau\geqslant 1$:

$$\tau(e^{|\beta|\varepsilon}-1) \leqslant 36e^{13}e^{2|\beta|\varepsilon}\sqrt{(\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}})\tau SAH\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8e\tau\big)+\log\big(8e\tau\big)^2\Big)}$$
$$+ 81e^{13}e^{2|\beta|\varepsilon}e^{|\beta|H}H^2SA\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8e\tau\big)+S\log\big(8e\tau\big)^2\Big)$$

Finally, we use lemma 31 with :

$$C = 36e^{13}\frac{\sqrt{(\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}})SAH}}{e^{|\beta|\varepsilon}-1} \quad ,A=\log(\frac{3SAH}{\delta}) \quad ,B=1$$
$$D=\frac{81e^{13}e^{\beta H}H^2SA}{e^{|\beta|\varepsilon}-1} \quad \text{and} \quad E=S$$

Which yield:

$$\tau \leqslant \frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}\frac{e^{2|\beta|\varepsilon}}{\big(e^{|\beta|\varepsilon}-1\big)^2}SAH\Big(\log(\frac{3SAH}{\delta})+1\Big)C_1^2+3\frac{e^{2|\beta|\varepsilon}}{e^{|\beta|\varepsilon}-1}e^{\beta H}H^2SA\Big(\log(\frac{3SAH}{\delta})+S\Big)C_1^2+1$$

Where $C_1=\frac{8}{5}\log\left(4e^{17}\frac{(S+1)(H+1)e^{|\beta H}SAH^2}{(e^{|\beta|\varepsilon}-1)}\right)$

In particular, assuming that ε is small enough that the first term dominates the second term then:

$$\tau \leqslant \frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}\frac{e^{2|\beta|\varepsilon}}{\big(e^{|\beta|\varepsilon}-1\big)^2}SAH\log(\frac{3SAH}{\delta})C_2^2$$

Where $C_2 = 3C_1$. We can finally hide the constants and the log terms to get:

$$\tau = \tilde{\mathcal{O}}\left(\frac{(e^{|\beta| G_{\max}(\mathcal{M})} - 1)^2}{e^{|\beta| G_{\max}(\mathcal{M})}} \frac{e^{2|\beta|\varepsilon}}{(e^{|\beta|\varepsilon} - 1)^2} SAH\right)$$

Finally to see that the algorithm is (ε, δ) PAC, At time τ:

$$\pi_1^\tau G_1(s_1) \leqslant (1 - e^{\beta\varepsilon}) \underset{\sim}{Z}_1^{\pi^\tau}$$

Since $Z_1^{\pi^\tau} \geqslant \underset{\sim}{Z}_1^{\pi^\tau}$, this is a stronger stopping condition than:

$$\pi_1^\tau G_1(s_1) \leqslant (1 - e^{\beta\varepsilon}) Z_1^{\pi^\tau}$$

Now, we write:

$$\left(V_1^\star - V_1^{\pi^\tau}\right)(s_1) = \frac{1}{\beta} \log\left(\frac{Z_1^\star}{Z_1^{\pi^\tau}}\right)(s_1) = \frac{1}{\beta} \log\left(1 + \frac{Z_1^\star - Z_1^{\pi^\tau}}{Z_1^{\pi^\tau}}\right)(s_1) \leqslant \frac{1}{\beta} \log\left(1 + \frac{\pi_1^\tau G_1(s)}{Z_1^{\pi^\tau}}\right)(s_1) \leqslant \varepsilon$$

■

Appendix C. Lower bound

We first state a change of measure for bandit models result from (Kaufmann et al., 2016):

Lemma 18 *Let $N_a(t) = \sum_{s=1}^t \mathbb{1}_{\{A_s = a\}}$ be the number of draws of arm a between the instants 1 and t and $N_a = N_a(\tau)$ be the total number of draws of arm a by some algorithm $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$.*

Let ν and ν' be two bandit models with K arms such that for all a, the distributions ν_a and ν'_a are mutually absolutely continuous. For any almost-surely finite stopping time σ with respect to $(\mathcal{F}_t)$,

$$\sum_{a=1}^K \mathbb{E}_\nu[N_a(\sigma)] \mathrm{KL}(\nu_a, \nu'_a) \geqslant \sup_{\mathcal{E} \in \mathcal{F}_\sigma} d(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E}))$$

where $d(x, y) = x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$ is the binary relative entropy, with the convention that $d(0, 0) = d(1, 1) = 0$.

We make the following assumption:

Assumption 1 Let $d = \lceil \log_A((S - 3)(A - 1) + 1) \rceil$. Assume that $H \geqslant 3d$

This assumption means that the horizon is long enough with respect to the size of MDP so that the agent can reach the reward state. We state the proof under the stronger assumption that $d = \log_A((S - 3)(A - 1) + 1)$ for simplicity. But as discussed in (Domingues et al., 2021), we can extend the construction to the general case by not having a full A-ary tree.

Condition A The bound is stated in the small ε regime. Let $c = e^{|\beta| H} - 1$, Assume that we have :

$$e^{|\beta|\varepsilon} < \begin{cases} \min\left\{\frac{4(c+1)^2}{2c^2+7c+4}, \frac{16}{13}\right\} & \beta > 0 \\ \min\left\{\frac{11c+8}{10c+8}, \frac{11}{10}\right\} & \beta < 0 \end{cases} \tag{21}$$

When $|\beta| H \geqslant \ln(2)$, the condition is reduced to the constants and gets harder to satisfy as β goes to 0. Condition 21 is used to have a valid construction in the proof below i.e to ensure the transition probabilities are in $[0, 1]$

Theorem 3 *Fix $S \geqslant 6, A \geqslant 2, H \in \mathbb{N}, \beta \in \mathbb{R}^\star$, and $\varepsilon, \delta \in (0,1)$ such that $\delta \leqslant 1/16$ and ε satisfy the condition (21). Then there exists an MDP $\mathcal{M}_0$ with S states, A actions, horizon H, and rewards in $[0,1]$ such that for every algorithm $\mathcal{A}$ output a policy $\hat{\pi}$ that is (ε, δ)-PAC for the entropic risk measure after sampling τ trajectories we have:*

$$\mathbb{E}_{\mathcal{M}_0}[\tau] \geqslant \frac{1}{1650} \frac{(e^{|\beta| G_{\max}(\mathcal{M}_0)} - 1)^2}{e^{|\beta| G_{\max}(\mathcal{M}_0)}} \frac{e^{2\min\{\beta,0\}\varepsilon} SAH}{(e^{|\beta|\varepsilon} - 1)^2} \log\left(\frac{1}{\delta}\right)$$

Proof Consider the following MDP defined in (Domingues et al., 2021)(with different transitions). We have three special states s_w(waiting state), s_g(the good absorbing state) and s_b(the bad absorbing state). The rest $S-3$ are arranged in the form of a full A-tree of depth $d-1$ denoted $\mathcal{L}$ whose root is s_{root}, this means that the number of states is

$$3 + \sum_{i=1}^{d-1} A^i = 3 + \frac{A^d - 1}{A - 1} = S$$

The action set is $\mathcal{A} = \{1, ..., A\}$ and let $a_w \in A$ be a fixed action, denote $L = A^{d-1}$. Let $\overline{H} \leqslant H - d$ be an integer to be chosen later.

The episode starts at s_w, for steps $h = 1, ..., \overline{H}$ we have the transition kernel:

$$p_h(s_w | a, s_w) = \mathbb{1}_{\{a = a_w, h \leqslant \overline{H}\}} \qquad \text{and} \qquad p_h(s_{\text{root}} | s_w, a) = 1 - p_h(s_w | s_w, a)$$

This means that for the first $\overline{H}$ steps, you can either chose the action a_w to stay in the waiting state s_w, or pick any other action and enter the tree at the next step, and at step $h = \overline{H}$ we exit regardless of the chosen action.

Once you enter the tree, the transition is deterministic. From any internal node x, the action a deterministically goes to the a-th child of x. Thus, after exactly $\overline{H} + d$ the policy reaches a leaf of the tree $\mathcal{L}$.

Define the index set of "triples"

$$\mathcal{U} = [\overline{H}] \times L \times [A], \quad |\mathcal{U}| = \overline{H} L A$$

A triple $u = (h, l, a)$ corresponds to:

- exiting at step h (so leaf step is $h + d$),
- reaching leaf l,
- choosing leaf action a

Fix two numbers $0 < p_- < p_+ < 1$.The baseline instance $\mathcal{M}_0$: for every triple $u = (t, l, a)$,

$$\mathbb{P}(s_g \mid t, l, a) = p_-$$

For $\mathcal{M}_0$, all leafs are the same and takes to the good state with probability p_- regardless of the chosen action. The Special instance $\mathcal{M}_u$ for each $u = (h^\star, l^\star, a^\star) \in \mathcal{U}$: identical to $\mathcal{M}_0$ except that at the single triple $u =$,

$$\mathbb{P}(s_g \mid u) = p_+$$

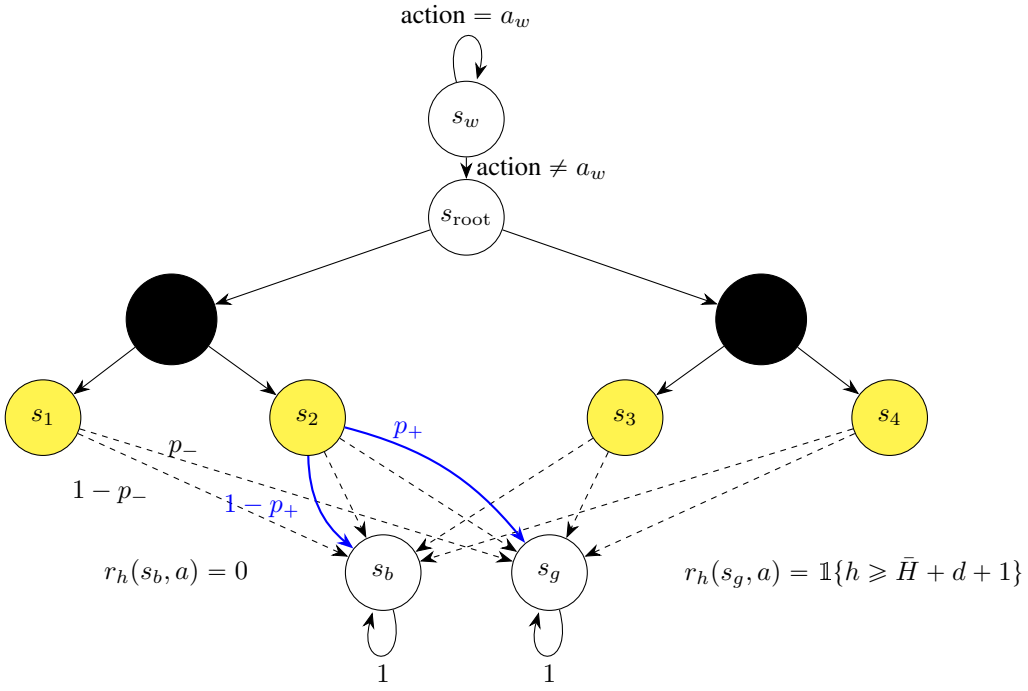

Figure 1: An example of the MDP construction. Here the state s_2 is optimal and has a better probability p_+ of landing in the good state s_g, p_+ and p_- are defined below in the proof. Figure reproduced from (Domingues et al., 2021)

and for all other triples $u' \neq u, \mathbb{P}(s_g \mid u') = p_-$. This means that there exists one unique optimal leaf $l^\star$ where the agent can chose one unique optimal action $a^\star$ when exiting precisely at step $h^\star$

Thus, the family is $\{\mathcal{M}_0\} \cup \{\mathcal{M}_u : u \in \mathcal{U}\}$, and any two instances differ in exactly one Bernoulli parameter at one triple.

Let

$$\tilde{H} = \overline{H} + d + 1, \quad H' = H - \overline{H} - d$$

Define rewards by

$$r_h(s, a) = \mathbf{1}\{s = s_g\}\mathbf{1}\{h \geqslant \tilde{H}\}$$

So rewards are in $[0, 1]$, and rewards only accumulate from $\tilde{H}$ onward. We give an illustration in figure (1) taken from (Domingues et al., 2021) and adapted to have our transitions that are a bit different from the original construction:

By construction, by time $\tilde{H}$ the chain has already entered s_g or s_b and is absorbing. Therefore the return is

$$G = \sum_{h=\tilde{H}}^{H} \mathbf{1}\{S_h = s_g\} = H'\mathbf{1}\{S_{\tilde{H}} = s_g\} \in \{0, H'\}$$

For any policy π and instance $\mathcal{M}$, let

$$p^{\mathcal{M}}(\pi) = \mathbb{P}_{\mathcal{M},\pi}(S_{\tilde{H}} = s_g)$$

The probability of being at the good state by the time $\tilde{H}$. By construction, by the time $\tilde{H}$ we are either in s_g or s_h, Then:

$$\mathbb{E}[e^{\beta G}] = (1 - p^M(\pi))1 + p^{\mathcal{M}}(\pi)e^{\beta H'} = 1 + cp^{\mathcal{M}}(\pi), \quad c = e^{\beta H'} - 1$$

Hence

$$V^{\mathcal{M}}(\pi) = \frac{1}{\beta}\log\left(1 + cp^{\mathcal{M}}(\pi)\right) \tag{22}$$

For a policy π, define the probability of success:

$$\nu_\pi(u) = \mathbb{P}_\pi(\text{the episode's exit/leaf/action triple equals } u), \quad u \in \mathcal{U}$$

Because each episode produces exactly one triple, we have

$$\sum_{u\in\mathcal{U}} \nu_\pi(u) = 1$$

In instance $\mathcal{M}_u$, since only when the realized triple equals the special one do we get probability p_+; otherwise p_-. The success probability is

$$p^{\mathcal{M}_u}(\pi) = p_- + (p_+ - p_-)\nu_\pi(u)$$

Let $\Delta = p_+ - p_- > 0$. The optimal policy in $\mathcal{M}_u$ can choose t, route to l, and pick action a deterministically so that $\nu_\pi(u) = 1$. Hence the optimal success probability is p_+, and

$$V^{\mathcal{M}_u,\star} = \frac{1}{\beta}\log(1 + cp_+)$$

Let us now chose p_+ and p_- so any ε-optimal policy must satisfy $\nu_\pi(u) > \frac{1}{2}$ in $\mathcal{M}_u$. The entropic criterion overweights rare high-return trajectories, so we make success(reaching the good state) rare by choosing $p_- \sim e^{-\beta H}$, more precisely:

$$p_- = \frac{1}{2(c+1)} \sim e^{-\beta H'}$$

We have:

$$V_\beta^{\mathcal{M}_u}(\pi) = \frac{1}{\beta}\log(1 + c(p_- + \Delta\nu_\pi(u)))$$

We chose Δ so that whenever we have a probability of success $\nu_\pi(u)$ smaller than $\frac{1}{2}$ the policy π is ε-suboptimal. When $\nu_\pi(u) \leqslant 1/2$, then $p^{\mathcal{M}_u}(\pi) \leqslant p_- + \Delta/2$. So it suffices to enforce

$$\frac{1}{\beta}\log\frac{1 + c(p_- + \Delta)}{1 + c(p_- + \Delta/2)} = \varepsilon \iff \frac{1 + c(p_- + \Delta)}{1 + c(p_- + \Delta/2)} = e^{\beta\varepsilon}$$

Hence, we must have:

$$\Delta = \frac{(3c+2)(e^{\beta\varepsilon} - 1)}{c(c+1)(2 - e^{\beta\varepsilon})}$$

Since we have that $e^{\beta\varepsilon} < \frac{4(c+1)^2}{2c^2+7c+4} < 2$ by condition 21, the construction is admissible in the sense that $0 < p_- < p_+ < 1$.

By this construction, we have for every $u \in \mathcal{U}$ and every policy π,

$$V^{\mathcal{M}_u}(\pi) \geqslant V^{\mathcal{M}_u,\star} - \varepsilon \implies \nu_\pi(u) > \frac{1}{2} \tag{23}$$

Indeed, if $\nu_\pi(u) \leqslant 1/2$, then $p^{\mathcal{M}_u}(\pi) \leqslant p_- + \Delta/2$, while the optimal policy achieves $p_+ = p_- + \Delta$. Hence

$$V^{\mathcal{M}_u,\star} - V^{\mathcal{M}_u}(\pi) \geqslant \frac{1}{\beta} \log \frac{1 + c(p_- + \Delta)}{1 + c(p_- + \Delta/2)} = \varepsilon$$

And the contraposition yields the claim (23). Let the algorithm output $\hat{\pi}$ at stopping time τ. Define the event

$$\mathcal{E}_u = \{\nu_{\hat{\pi}}(u) > 1/2\}$$

By the previous remark and (ε, δ)-PAC correctness,

$$\mathbb{P}_u(\mathcal{E}_u) \geqslant 1 - \delta \quad \forall u \in \mathcal{U}$$

Also, since $\sum_u \nu_{\hat{\pi}}(u) = 1$, at most one u can satisfy $\nu_{\hat{\pi}}(u) > 1/2$. Hence the events $\{\mathcal{E}_u\}_{u \in \mathcal{U}}$ are mutually exclusive and in particular under $\mathcal{M}_0$,

$$\sum_{u \in \mathcal{U}} \mathbb{P}_0(\mathcal{E}_u) \leqslant 1$$

Let $N_u(\tau)$ be the number of episodes $k \leqslant \tau$ in which the algorithm's realized triple equals u. Then

$$\sum_{u \in \mathcal{U}} N_u(\tau) = \tau \quad \text{a.s.}$$

Fix $u \in \mathcal{U}$. We apply lemma 18 for the event $\mathcal{E}_u$, the two instances $\mathcal{M}_u$ and $\mathcal{M}_0$ differ only in the Bernoulli transition at triplet u hence:

$$\begin{aligned}\mathbb{E}_0[N_u(\tau)] d(p_-, p_+) &\geqslant d(\mathbb{P}_0(\mathcal{E}_u), \mathbb{P}_u(\mathcal{E}_u)) \geqslant (1 - \mathbb{P}_0(\mathcal{E}_u)) \log\left(\frac{1}{1 - \mathbb{P}_u(\mathcal{E}_u))}\right) - \log(2) \\ &\geqslant (1 - \mathbb{P}_0(\mathcal{E}_u)) \log\left(\frac{1}{\delta}\right) - \log(2)\end{aligned}$$

Where we used the (ε, δ)-PAC in the last inequality. We sum over $u \in \mathcal{U}$:

$$\mathbb{E}_0[\tau] = \sum_{u \in \mathcal{U}} \mathbb{E}_0[N_u(\tau)] \geqslant \frac{1}{d(p_-, p_+)} \left(\sum_{u \in \mathcal{U}} (1 - \mathbb{P}_0(\mathcal{E}_u)) \log\left(\frac{1}{\delta}\right) - \log(2) \right) \geqslant \frac{1}{2d(p_-, p_+)} |\mathcal{U}| \log\left(\frac{1}{\delta}\right)$$

Where we used in the final inequality that $\sum_{u \in \mathcal{U}} \mathbb{P}_0(\mathcal{E}_u) \leqslant 1$ and the condition $\delta \leqslant \frac{1}{16}$ and since:

$$d(p_-, p_+) \leqslant \frac{(p_+ - p_-)^2}{p_-(1 - p_-)} \leqslant (c+1) \frac{(e^{\beta\varepsilon} - 1)^2 \left(\frac{3c+2}{2(c+1)}\right)^2}{c^2 (1 - \frac{1}{2} e^{\beta\varepsilon})^2} = \frac{(c+1)}{c^2} \frac{(e^{\beta\varepsilon} - 1)^2}{(1 - \frac{1}{2} e^{\beta\varepsilon})^2} \left(\frac{3c+2}{2(c+1)}\right)^2$$

Using the condition $e^{\beta\varepsilon} \leqslant \frac{16}{13}$ we get:

$$d(p_-, p_+) \leqslant \frac{1521}{100} \frac{(c+1)}{c^2} (e^{\beta\varepsilon} - 1)^2$$

Hence:

$$\mathbb{E}_0[\tau] \geqslant \frac{50}{1521}\frac{c^2}{c+1}\frac{\big(H-\overline{H}-d\big)LA}{(e^{\beta\varepsilon}-1)^2}\log\left(\frac{1}{\delta}\right)$$

Since the number of leaves is given by $L = (1-1/A)(S-3)+1/A \geqslant S/4$ (for $A \geqslant 2, S \geqslant 6$), and taking $\overline{H} = H/3$ with $d \leqslant H/3$, we obtain the sample complexity lower bound:

$$\mathbb{E}_0[\tau] \geqslant \frac{1}{1650}\frac{\left(e^{\beta(H-\overline{H}-1)}-1\right)^2}{e^{\beta(H-\overline{H}-1)}}\frac{SAH}{(e^{\beta\varepsilon}-1)^2}\log\left(\frac{1}{\delta}\right)$$

Since in the constructed MDP, we only accumulate rewards after $\tilde{H}$ and the optimal policy accumulate rewards across all subsequent steps, we have: $G_{\max}(\mathcal{M}_0) = H - \tilde{H} + 1 = H - \overline{H} - d$ Hence:

$$\mathbb{E}_0[\tau] \geqslant \frac{1}{1650}\frac{\left(e^{\beta G_{\max}(\mathcal{M}_0)}-1\right)^2}{e^{\beta G_{\max}(\mathcal{M}_0)}}\frac{SAH}{(e^{\beta\varepsilon}-1)^2}\log\left(\frac{1}{\delta}\right)$$

For $\beta > 0$ (risk-seeking entropic criterion), we choose p_- so that transitioning to the good absorbing state is a rare event. This makes upside tail events hard to detect and estimate, and since the entropic objective overweights favorable rare outcomes, this again leads to larger sample complexity.

$$\begin{aligned}\mathbb{E}[e^{\beta G}] &= (1-p^{\mathcal{M}}(\pi))1 + p^{\mathcal{M}}(\pi)e^{\beta H'} = e^{\beta H'}\left((1-p^{\mathcal{M}}(\pi))e^{|\beta|H'} + p^{\mathcal{M}}(\pi)\right)\\ &= e^{\beta H'}\left((1-p^{\mathcal{M}}(\pi))(e^{|\beta|H'}-1)+1\right) = e^{\beta H'}\left((1-p^{\mathcal{M}}(\pi))c+1\right)\end{aligned}$$

And we have the entropic risk measure:

$$V^{\mathcal{M}}(\pi) = -\frac{1}{|\beta|}\log\left(e^{\beta H'}\left((1-p^{\mathcal{M}}(\pi))c+1\right)\right) = H' - \frac{1}{|\beta|}\log\left((1-p^{\mathcal{M}}(\pi))c+1\right)$$

Where $c = e^{|\beta|H'}-1$.

For $\beta < 0$, the entropic criterion is especially sensitive to adverse tail events, so we instead make failure rare by choosing $p_- \sim 1 - e^{-|\beta|H}$, more precisely:

$$p_- = 1 - \frac{1}{2(c+1)} \sim 1 - e^{-\beta H'}$$

We derive p_+ similarly to $\beta > 0$ and we find if $p_+ = p_- + \Delta$ then:

$$\Delta = \frac{(3c+2)(e^{|\beta|\varepsilon}-1)}{c(c+1)(e^{|\beta|\varepsilon}-\frac{1}{2})}$$

Again, since we have $e^{|\beta|\varepsilon} < \frac{11c+8}{10c+8}$, this construction is admissible in the sense that $0 < p_- < p_+ < 1$.

And we prove in the same way that: for every $u \in \mathcal{U}$ and every policy π,

$$V^{\mathcal{M}_u}(\pi) \geqslant V^{M_u,\star} - \varepsilon \implies \nu_\pi(u) > \frac{1}{2}$$

The rest of the argument (change of measure and summing over $\mathcal{U}$ goes the same way)and we finally upper bound the kl divergence:

$$d(p_-, p_+) \leqslant (c+1)\frac{(e^{|\beta|\varepsilon}-1)^2\left(\frac{3c+2}{2(c+1)}\right)^2}{c^2(e^{|\beta|\varepsilon}-\frac{1}{2})^2} \leqslant \frac{9(c+1)}{c^2}\frac{(e^{|\beta|\varepsilon}-1)^2}{e^{2\beta\varepsilon}}\left(\frac{3c+2}{2(c+1)}\right)^2$$

Hence by having a looser constant to match the $\beta > 0$ lower bound:

$$\mathbb{E}_0[\tau] \geqslant \frac{1}{1650}\frac{\left(e^{|\beta|G_{\max}(\mathcal{M}_0)}-1\right)^2}{e^{|\beta|G_{\max}(\mathcal{M}_0)}}\frac{e^{2\beta\varepsilon}SAH}{(e^{|\beta|\varepsilon}-1)^2}\log\left(\frac{1}{\delta}\right)$$

■

Appendix D. Regret bounds

First we prove theorem 5. We only give the proof for $\beta > 0$ but the proof for $\beta < 0$ is similar to the sample complexity analysis

Proof Consider algorithm 1 run indefinitely with π^{t+1} the algorithm computed from n^t and $\widehat{p}^t$.Define the accumulated regret as:

$$R(K) = \sum_{t=1}^{K-1} V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1)$$

We will work on the good event $\mathcal{E}$. We have:

$$\begin{aligned}
R(K) &= \sum_{t=1}^{K-1} V_1^\star(s_1) - V_1^{\pi_{t+1}}(s_1) \\
&= \sum_{t=1}^{K-1} \frac{1}{\beta}\log\left(\frac{Z_1^\star}{Z_1^{\pi^{t+1}}}\right)(s_1) \\
&= \sum_{t=1}^{K-1} \frac{1}{\beta}\log\left(1+\frac{Z_1^\star - Z_1^{\pi^{t+1}}}{Z_1^{\pi^{t+1}}}\right)(s_1) \\
&\leqslant \sum_{t=1}^{K-1} \frac{1}{\beta}\log\left(1+\frac{\pi_1^{t+1}G_1^t(s_1)}{Z_1^{\pi^{t+1}}}\right)(s_1) \\
&\leqslant \sum_{t=1}^{K-1} \frac{1}{\beta}\frac{\pi_1^{t+1}G_1^t(s_1)}{Z_1^{\pi^{t+1}}}(s_1)
\end{aligned}$$

The inequality follows since it's at each step we have

$$Z_1^\star - Z_1^{\pi^{t+1}}(s_1) \leqslant \pi_1^{t+1}G_1^t(s_1)$$

By following the same analysis as the paper we find that for any episode t we have:

$$(\pi_1 G_1^t)(s_1) \leqslant 36e^{13}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}}\sqrt{\sum_{h=1}^H\sum_{s,a} p_h^{t+1}(s,a)\left(\frac{\alpha^\star\left(n_h^t(s,a)\right)}{n_h^t(s,a)}\wedge 1\right)}$$
$$+ 84e^{13}e^{\beta H}\sum_{h=1}^H\sum_{s,a} p_h^{t+1}(s,a)\left(\frac{\alpha\left(n_h^t(s,a)\right)}{n_h^t(s,a)}\wedge 1\right)$$

Now to compute the regret we sum up to the K episodes and using Cauchy-Schwartz inequality:

$$\begin{aligned}
R(K) &\leqslant 36e^{13}\sum_{t=1}^{K-1}\frac{1}{\beta}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}\frac{1}{\beta}e^{\beta H}\sum_{t=1}^{K-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big) \\
&\leqslant 36e^{13}\frac{1}{\beta}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}}\sqrt{K}\sqrt{\sum_{t=1}^{K-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha^\star\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)} \\
&\quad + 84e^{13}\frac{1}{\beta}e^{\beta H}\sum_{t=1}^{K-1}\sum_{h=1}^{H}\sum_{s,a}p_h^{t+1}(s,a)\big(\frac{\alpha\big(n_h^t(s,a)\big)}{n_h^t(s,a)}\wedge 1\big)
\end{aligned}$$

Then using the pseudo-counts lemma we have:

$$\begin{aligned}
R(K) &\leqslant 36e^{13}\frac{1}{\beta}\sqrt{\frac{(e^{\beta G_{\max}(\mathcal{M})}-1)^2}{e^{\beta G_{\max}(\mathcal{M})}}KSAH\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8eK\big)+\log\big(8eK\big)^2\Big)} \\
&\quad + 84e^{13}\frac{e^{\beta H}}{\beta}SAH\Big(\log\big(\frac{3SAH}{\delta}\big)\log\big(8eK\big)+S\log\big(8eK\big)^2\Big)
\end{aligned}$$

Hence omitting constants we have:

$$R(K)=\mathcal{O}\left(\frac{1}{|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}(\mathcal{M})}-1)^2}{e^{|\beta|G_{\max}(\mathcal{M})}}KSAH\log\left(\frac{SAHK}{\delta}\right)}+\frac{e^{|\beta|H}}{\beta}S^2AH\log^2\left(\frac{SAHK}{\delta}\right)\right).$$

■

Now we prove the lower bound, that we state first as a theorem:

Theorem 20 *Fix $\beta\neq 0$. Let $\mathcal{M}_\varepsilon$ be the hard family used in the BPI lower bound and denote by $G_{\max}$ the maximal achievable return of the instances in this family (which does not depend on ε). Then, for every regret algorithm $\mathcal{A}$ and every $K\geqslant 2e^{|\beta|G_{\max}}SAH$ we have: there exists an instance $M\in\mathfrak{M}_{\varepsilon_K}$(for a well chosen ε_K) and a constant $\eta>0$ such that*

$$\mathbb{E}_M[R_M(K)]\geqslant\frac{\eta}{6400|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}}SAHK}$$

Consequently

$$\inf_{\mathcal{A}}\sup_{M}\mathbb{E}_M[R_M(K)]\geqslant\frac{\eta}{32|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}}SAHK}.$$

Proof Fix an arbitrary regret algorithm $\mathcal{A}$. When $\mathcal{A}$ is run for K episodes on an instance M, let $\pi^1,\pi^2,\ldots,\pi^K$ be the policies played by the algorithm. The cumulative regret is

$$R_M(K)=\sum_{t=1}^{K}\Big(V_1^{\star,M}(s_1)-V_1^{\pi^t,M}(s_1)\Big).$$

Choose

$$\varepsilon_K = \frac{\eta}{|\beta|}\sqrt{\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}K}SAH}$$

Where η is a constant such that: $\eta \leqslant \sqrt{2}$ and $\eta^2 \leqslant \frac{c_{\text{BPI}}e^{-2}\log(32)}{8}$ (with $c_{\text{BPI}} = \frac{1}{1650}$). The condition on K allow us to apply the lower bound on the BPI with $\varepsilon = \varepsilon_K$. Now suppose for contradiction that: Now suppose, toward a contradiction, that

$$\sup_{M\in\mathfrak{M}_{\varepsilon_K}} \mathbb{E}_M[R_M(K)] < \delta_0 K\varepsilon_K$$

We construct a fixed-budget BPI algorithm from $\mathcal{A}$: Run $\mathcal{A}$ for exactly K episodes. After these K episodes, sample $J \sim \text{Unif}\{1,\dots,K\}$ independently of the interaction history, and output $\widehat{\pi} = \pi^J$.

For any $M \in \mathfrak{M}_{\varepsilon_K}$, by independence of J and linearity of expectation,

$$\begin{aligned}\mathbb{E}_M\left[V_1^{\star,M}(s_1) - V_1^{\widehat{\pi},M}(s_1)\right] &= \frac{1}{K}\sum_{t=1}^{K}\mathbb{E}_M\left[V_1^{\star,M}(s_1) - V_1^{\pi^t,M}(s_1)\right]\\ &= \frac{\mathbb{E}_M[R_M(K)]}{K}\\ &< \delta_0\varepsilon_K\end{aligned}$$

Since $V_1^{\star,M}(s_1) - V_1^{\widehat{\pi},M}(s_1) \geqslant 0$ Markov inequality gives

$$\mathbb{P}_M\left(V_1^{\star,M}(s_1) - V_1^{\widehat{\pi},M}(s_1) > \varepsilon_K\right) \leqslant \frac{\mathbb{E}_M[V_1^{\star,M}(s_1) - V_1^{\widehat{\pi},M}(s_1)]}{\varepsilon_K} < \delta_0$$

Therefore,

$$\mathbb{P}_M\left(V_1^{\star,M}(s_1) - V_1^{\widehat{\pi},M}(s_1) \leqslant \varepsilon_K\right) \geqslant 1-\delta_0$$

We will show now this contradicts the BPI lower bound. The running time is deterministic and satisfies $\mathbb{E}_{\mathcal{M}_0}[\tau] = K$ We now show that this contradicts the BPI lower bound. Since $|\beta|\varepsilon_K \leqslant 1$ by the condition on K and η, we have

$$e^{2\min\{\beta,0\}\varepsilon_K} \geqslant e^{-2|\beta|\varepsilon_K} \geqslant e^{-2}$$

Also, using the inequality $e^x - 1 \leqslant 2x$ for $x \in [0,1]$, we obtain

$$e^{|\beta|\varepsilon_K} - 1 \leqslant 2|\beta|\varepsilon_K$$

Applying the BPI lower bound with $\delta_0 = 1/32$, we get

$$\begin{aligned}T_{\text{BPI}}(\varepsilon_K,\delta_0) &\geqslant c_{\text{BPI}}\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}}\frac{e^{2\min\{\beta,0\}\varepsilon_K}SAH}{(e^{|\beta|\varepsilon_K}-1)^2}\log(32)\\ &\geqslant c_{\text{BPI}}\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}}\frac{e^{-2}SAH}{4\beta^2\varepsilon_K^2}\log(32)\end{aligned}$$

By definition of ε_K,

$$\beta^2\varepsilon_K^2 = \eta^2\frac{(e^{|\beta|G_{\max}}-1)^2}{e^{|\beta|G_{\max}}}\frac{SAH}{K}$$

Substituting this into the previous display yields

$$T_{\text{BPI}}(\varepsilon_K, \delta_0) \geqslant \frac{c_{\text{BPI}} e^{-2} \log(32)}{4\eta^2} K$$

Since by the definition of η we have:

$$\frac{c_{\text{BPI}} e^{-2} \log(32)}{4\eta^2} \geqslant 2$$

Thus

$$K = \mathbb{E}_{\mathcal{M}_0}[\tau] \geqslant T_{\text{BPI}}(\varepsilon_K, \delta_0) \geqslant 2K > K$$

This contradicts the existence of the fixed-budget $(\varepsilon_K, \delta_0)$-PAC BPI algorithm using exactly K episodes. Therefore, the contradiction assumption must be false. Hence

$$\sup_{M \in \mathfrak{M}_{\varepsilon_K}} \mathbb{E}_M[R_M(K)] \geqslant \delta_0 K \varepsilon_K$$

Substituting the definition of ε_K, we get

$$\begin{aligned}\sup_{M \in \mathfrak{M}_{\varepsilon_K}} \mathbb{E}_M[R_M(K)] &\geqslant \delta_0 K \frac{\eta}{|\beta|} \sqrt{\frac{(e^{|\beta| G_{\max}} - 1)^2}{e^{|\beta| G_{\max}}} \frac{SAH}{K}} \\ &= \frac{\eta}{32|\beta|} \sqrt{\frac{(e^{|\beta| G_{\max}} - 1)^2}{e^{|\beta| G_{\max}}} SAHK}\end{aligned}$$

Finally, since $\mathfrak{M}_{\varepsilon_K}$ is a subset of the full MDP class under consideration, and since $\mathcal{A}$ was arbitrary,

$$\inf_{\mathcal{A}} \sup_{M} \mathbb{E}_M[R_M(K)] \geqslant \frac{\eta}{32|\beta|} \sqrt{\frac{(e^{|\beta| G_{\max}} - 1)^2}{e^{|\beta| G_{\max}}} SAHK}$$

■

Appendix E. Experiments

For our experiments, we consider a toy MDP that consists of 8 states and two actions, **safe** and **risky**. Starting from s_0, **safe** deterministically walks along the bridge $s_0 \to s_1 \to s_2 \to s_3 \to s_4 \to s_5 \to s_g$, where s_g is an absorbing goal state. The **risky** action attempts shortcuts: from early states it jumps directly to s_4 but can fall into an absorbing bad state s_b with state-dependent probability (start/mid/final risk); from s_4 it makes a final dash to s_g that can also fail and transition to s_b. Both s_g and s_b are absorbing and we receive reward 1 in s_g for the rest of the horizon. For analogy, think of it as a child standing at the top of a long staircase. Taking safe means walking down (or along) the stairs one step at a time, steadily progressing. Taking risky means you try to jump over several steps at once to land much farther ahead saving time, but with some chance of missing the landing and falling into a pit (failure). After the staircase you reach a narrow bridge: safe is crossing it normally, while risky is a final dash across the bridge, which is faster but again risks falling into the pit. If you make it, you end in the goal; if you fall, you're stuck in failure. The MDP is visualized in the figure 2. In our experiments we take $\left\{p_{\text{risk,start}}, p_{\text{risk,mid}}, p_{\text{risk,final}}, p_{\text{fall,dash}}\right\} = \left\{0.95, 0.75, 0.25, 0.85\right\}$, we also take $\varepsilon = 0.2$ and $\delta = 0.1$

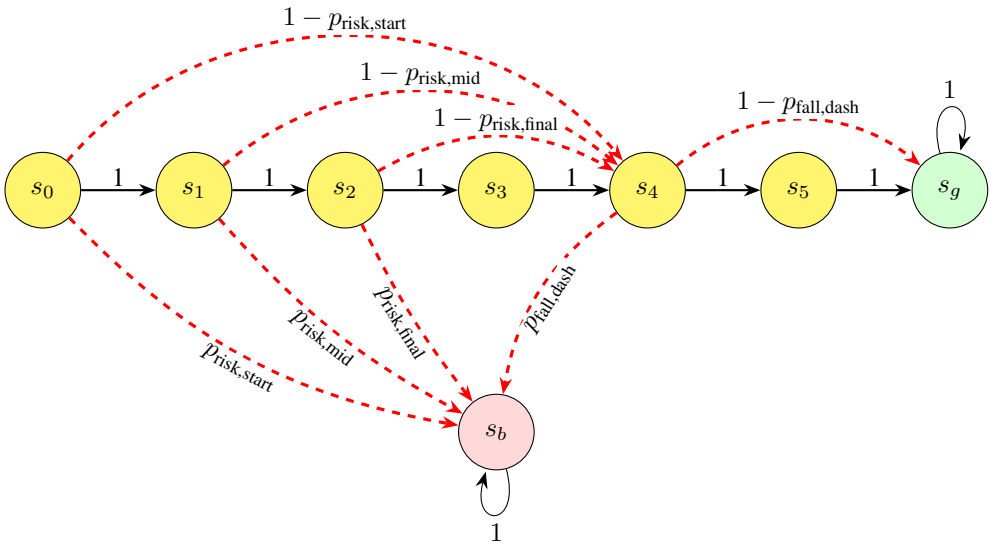

Figure 2: toy MDP. Safe action follows the straight chain to s_g; risky action can shortcut but may fall to s_b. **safe** action is represented by a black edge and **risky** action by a dashed red edge.

E.1. Sample complexity

We study the sample complexity of our method as a function of the risk sensitivity β and the horizon H. In the first sweep, we fix $H = 7$ and run the algorithm for $\beta \in \{2.5, 3.0, 3.5, 4.0, 4.5\}$ until the stopping condition is met. In the second sweep, we fix $\beta = 0.5$ and vary the horizon $H \in \{5, 7, 9, 11, 13, 15, 17\}$. For each configuration, we run 15 independent trials (different random seeds) and plot the mean stopping time τ; the y-axis is shown on a log scale. For Figure 4, the

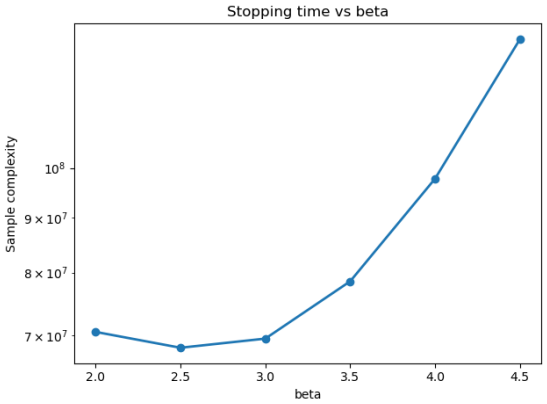

Figure 3: Sample complexity in function of β (log-scale y-axis)

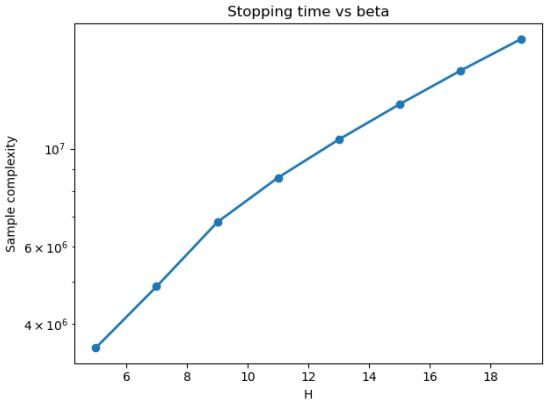

Figure 4: Sample complexity in function of H (log-scale y-axis)

stopping time τ increases sharply as β grows. Moreover, the approximately linear trend on the log-scale y-axis suggests that $\log \tau$ increases roughly linearly with β, i.e., the sample complexity is consistent with an exponential scaling $\tau \approx \exp(c\,\beta)$ over the tested range. This is in line with

the intuition behind the entropic objective, where larger β amplifies tail events and makes BPI more demanding.

Interestingly, the effective slope c is not constant across β. We hypothesize that this variation is driven by the χ^2-divergence term $\chi^2(\mathbb{P}^\pi_\beta, \mathbb{P}^\pi)$ discussed after equation 9, which depends on the policy induced at each β. In our sweep, the learned policy changes substantially as β increases, reflecting policy changes as the agent shows increasingly risk-seeking behavior, and then stabilizes around $\beta \approx 3.5$. Beyond this point, further increases in β do not change the policy, which may explain the change in scaling behavior. For Figure 4, we observe a similarly sharp increase in the stopping time τ as the horizon H grows and is roughly linear on the log scale

E.2. Regret bounds

We now compare our approach to other regret algorithms. At each episode, we evaluate the current policy returned by each algorithm from the initial state s_0. We then compute the cumulative regret for each algorithm defined by:

$$R(K) = \sum_{k=1}^{K} V^\star(s_0) - V^{\pi^{k+1}}(s_0)$$

Where π^{k+1} is the policy returned by the algorithm at the end of episode k. We cap the stopping time at 10^7 episodes and plot the regret for $\beta \in [1.0, 2.0, 3.0, 4.0]$. We compare our method to multiple regret algorithms:

- RSVI and RSQ from (Fei et al., 2020)
- RSVI2 and RSQ2 from (Fei et al., 2021)
- UCB ADVANTAGE from (Hu and Leung, 2023)
- RODI (OTP and PTO variants) and ROVI from (Liang and Luo, 2024)

We report the findings of our experiment:

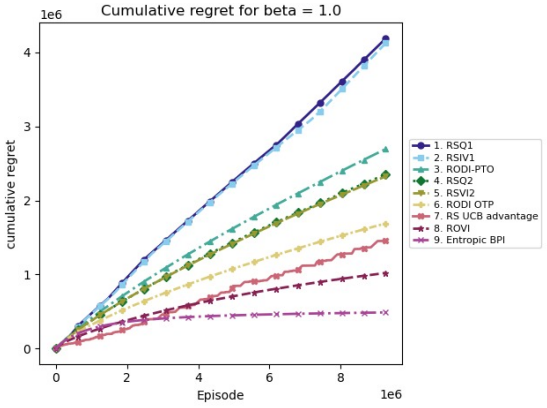

Figure 5: Regret for different algorithms for $\beta = 1.0$

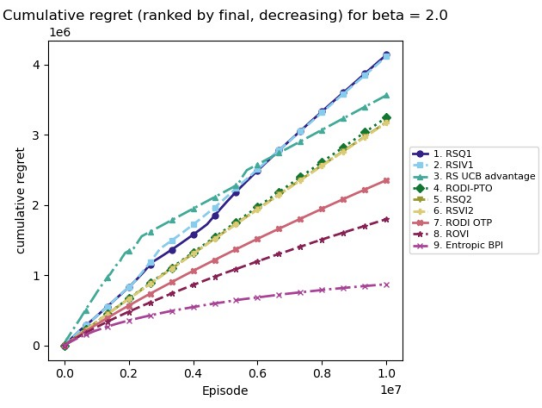

Figure 6: Regret for different algorithms for $\beta = 2.0$

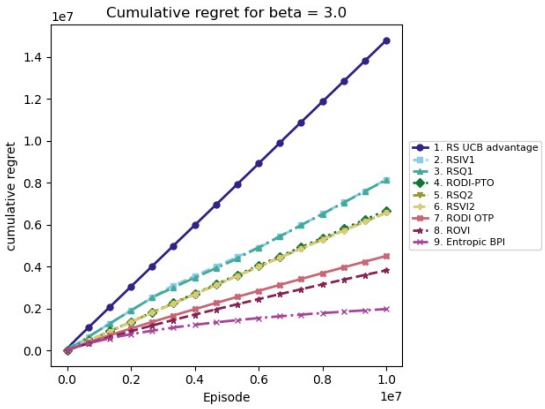

Figure 7: Regret for different algorithms for $\beta = 3.0$

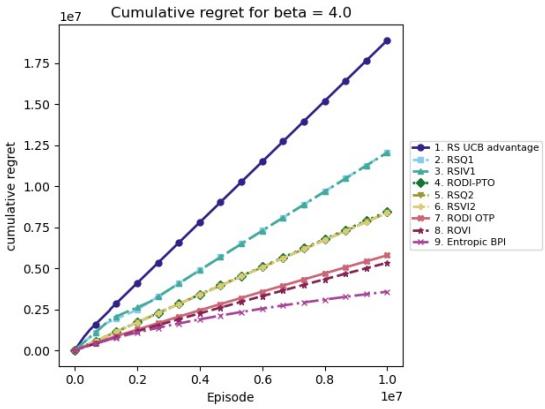

Figure 8: Regret for different algorithms for $\beta = 4.0$

Figures 5, 6, 7, 8 show that our algorithm consistently achieves the lowest cumulative-regret trajectory over all values of β. Since cumulative regret is computed by evaluating the current policy from the initial state each episode and summing the resulting value gap to optimality, the consistently flatter Entropic BPI curve suggests stronger learning efficiency and performance.

Appendix F. Concentration inequalities

F.1. Sanov's theorem

First we introduce the K divergence concentration inequality derived via Sanov's theorem

Lemma 21 (High-probability KL bound via Sanov) *Let $\Sigma_m = \{q \in \mathbb{R}^m_{\geqslant 0} : \sum_{i=1}^m q_i = 1\}$ and let $p \in \Sigma_m$. Let $X_1, \ldots, X_n$ be i.i.d. with law p, and let the empirical distribution $\widehat{p}_n$ be*

$$\widehat{p}_n(i) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{X_k = i\}, \qquad i \in [m].$$

Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$D(\widehat{p}_n \| p) \leqslant \frac{(m-1)\log(n+1) + \log(1/\delta)}{n}.$$

Proof Let $\Sigma_m = \left\{q \in \mathbb{R}^m_{\geqslant 0} : \sum_{i=1}^m q_i = 1\right\}$ be the $(m-1)$-simplex and let $p \in \Sigma_m$. Let $X_1, \ldots, X_n$ be i.i.d. with law p, and let $\widehat{p}_n$ denote the empirical distribution by

$$\widehat{p}_n(i) := \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}\{X_k = i\}, \qquad i \in [m]$$

Sanov's theorem (Theorem 11.4.1 Cover and Thomas (2006)) states that for any set $E \subseteq \Sigma_m$

$$\Pr\left(\widehat{p}_n \in E\right) \leqslant (n+1)^{(m-1)} \exp\Big(-n \inf_{q \in E} D(q\|p)\Big)$$

Fix $\varepsilon > 0$ and take the set $E = \{q \in \Sigma_m | D(p||q) \geqslant \varepsilon\}$ then $\inf_{q\in E} D(q\|p) = \varepsilon$ and by Sanov's theorem:

$$\Pr\left(D(\widehat{p}_n||p) \geqslant \varepsilon\right) \leqslant (n+1)^m \exp\Big(-n\varepsilon\Big)$$

We turn it to high probability bound by setting the r.h.s to δ which yield:

$$\varepsilon = \frac{m\log(n+1) + \log(1/\delta)}{n}$$

Doing a union bound on all state action pairs and all time steps we get that with probability $1-\delta$ we have:

$$D(\widehat{p}_n||p) \leqslant \frac{(m-1)\log(n+1) + \log(SAH/\delta)}{n}$$

■

F.2. Concentration inequality for Bernoulli random variables

We state a deviation inequality for Bernoulli random variables from Dann et al. (2017)(Lemma F.4):

Lemma 22 *Let $\mathcal{F}_i$ for $i = 1\ldots$ be a filtration and $X_1, \ldots X_n$ be a sequence of Bernoulli random variables with $\mathbb{P}(X_i = 1|\mathcal{F}_{i-1}) = P_i$ with P_i being $\mathcal{F}_{i-1}$-measurable and X_i being $\mathcal{F}_i$ measurable. It holds that*

$$\mathbb{P}\left(\exists n : \sum_{t=1}^{n} X_t < \sum_{t=1}^{n} P_t/2 - \log(\frac{1}{\delta})\right) \leqslant \delta$$

F.3. Self-normalized Bernstein inequality

We state the self-normalized Bernstein-type inequality by (Menard et al., 2021).

Lemma 23 *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 $|Y_t| \leqslant b$ 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. Then For all $\delta > 0$:

$$\mathbb{P}\left(\exists t \geqslant 1, (V_t/b^2 + 1)h\left(\frac{b|S_t|}{V_t + b^2}\right) \geqslant \log(1/\delta) + \log(4e(2t+1))\right) \leqslant \delta.$$

The previous inequality can be weakened to obtain a more explicit bound: with probability at least $1-\delta$, for all $t \geqslant 1$,

$$|S_t| \leqslant \sqrt{2V_t \log(4e(2t+1)/\delta)} + 3b\log(4e(2t+1)/\delta).$$

F.4. KL-Bernstein inequality

We state a KL-Bernstein inequality from (Talebi and Maillard, 2018):

Lemma 24 *Let $p, q \in \Sigma_S$, where Σ_S denotes the probability simplex of dimension $S-1$. For all $\alpha > 0$, for all functions f defined on $\mathcal{S}$ with $0 \leqslant f(s) \leqslant b$, for all $s \in \mathcal{S}$, if $\mathrm{KL}(p,q) \leqslant \alpha$ then*

$$|pf - qf| \leqslant \sqrt{2\mathrm{Var}_q(f)\alpha} + \frac{2}{3}b\alpha$$

where we use the expectation operator defined as $pf \triangleq \mathbb{E}_{s\sim p} f(s)$ and the variance operator defined as $\mathrm{Var}_p(f) \triangleq \mathbb{E}_{s\sim p}(f(s) - \mathbb{E}_{s'\sim p} f(s'))^2 = p(f-pf)^2$.

Appendix G. Technical results

The next lemma introduces the entropic variance under a fixed policy π, defined as the conditional variance of the exponentiated return $e^{\beta G_h}$ around its entropic value $e^{\beta Q_h^\pi}$. It shows that σQ_h^π satisfies a Bellman-style recursion: for each step h and state–action pair (s,a), $\sigma Q_h^\pi(s,a)$ decomposes into (i) the variance under the next-state transition $p_h(\cdot \mid s,a)$ of $e^{\beta(r_h(s,a)+V_{h+1}^\pi(s'))}$ and (ii) the expected next-step entropic variance scaled by $e^{2\beta r_h(s,a)}$ with terminal condition $\sigma_{H+1}^\pi V_h^\pi = 0$.

Lemma 25 (Entropic variance recursion) *Fix a (deterministic) policy π.*

$$\sigma Q_h^\pi(s,a) = \mathbb{E}^\pi\left[\left(e^{\beta R_h^\pi(s,a)} - e^{\beta Q_h^\pi(s,a)}\right)^2 \;\middle|\; S_h = s, A_h = a\right]$$

with terminal condition $\sigma_{H+1}^\pi(s,a) = 0$ and:

$$\sigma V_h^\pi(s) = \sigma Q_h^\pi(s,\pi(s))$$

Then, for every $h \in \{1,\dots,H\}$ and $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\sigma Q_h^\pi(s,a) = e^{2\beta r_h(s,a)}\,\mathrm{Var}_{S'\sim p_h(\cdot|s,a)}\left(Z_{h+1}^\pi(S')\right) + e^{2\beta r_h(s,a)}\,(p_h\,\sigma V_{h+1}^\pi)(s,a) \tag{24}$$

Proof Fix $h \in \{1,\dots,H\}$ and (s,a). By definition of $U_h^\pi = e^{\beta Q_h^\pi}$ (exponential entropic Q), we have

$$U_h^\pi(s,a) = \mathbb{E}^\pi\left[e^{\beta R_h^\pi} \mid S_h = s,\ A_h = a\right],$$

hence $\sigma Q_h^\pi(s,a) = \mathrm{Var}\left(e^{\beta R_h^\pi} \mid S_h = s, A_h = a\right)$.

Since $r_h(s,a)$ is deterministic given $(s_h,a_h) = (s,a)$,

$$e^{\beta R_h^\pi} = e^{\beta r_h(s,a)}\,e^{\beta R_{h+1}^\pi} \quad\Rightarrow\quad \sigma Q_h^\pi(s,a) = e^{2\beta r_h(s,a)}\,\mathrm{Var}\left(e^{\beta R_{h+1}^\pi} \mid s_h = s,\ a_h = a\right)$$

Apply the law of total variance w.r.t. S_{h+1}:

$$\begin{aligned}\mathrm{Var}\left(e^{\beta R_{h+1}^\pi} \mid s_h = s,\ a_h = a\right) = \mathbb{E}\Big[\mathrm{Var}\left(e^{\beta R_{h+1}^\pi} \mid S_{h+1} = s_{h+1}\right) \mid S_h = s,\ A_h = a\Big] \\ + \mathrm{Var}\left(\mathbb{E}\left[e^{\beta R_{h+1}^\pi} \mid S_{h+1} = s_{h+1}\right] \mid S_h = s,\ A_h = a\right)\end{aligned}$$

For the first term, conditioning on $S_{h+1} = s'$ and following π thereafter gives

$$\mathrm{Var}\left(e^{\beta R^{\pi}_{h+1}} \mid S_{h+1} = s'\right) = \sigma V^{\pi}_{h+1}(s')$$

so

$$\mathbb{E}\left[\mathrm{Var}\left(e^{\beta R^{\pi}_{h+1}} \mid S_{h+1}\right) \mid S_h = s,\ A_h = a\right] = \mathbb{E}_{S'\sim p_h(\cdot|s,a)}\left[\sigma V^{\pi}_{h+1}(S')\right] = (p_h\,\sigma V^{\pi}_{h+1})(s,a)$$

For the second term, by definition of $Z^{\pi}_{h+1}(s') = e^{\beta V^{\pi}_{h+1}(s')} = \mathbb{E}^{\pi}[e^{\beta G_{h+1}} \mid s_{h+1} = s']$,

$$\mathrm{Var}\left(\mathbb{E}\left[e^{\beta G_{h+1}} \mid S_{h+1}\right] \mid S_h = s,\ A_h = a\right) = \mathrm{Var}_{S'\sim p_h(\cdot|s,a)}\left(Z^{\pi}_{h+1}(S')\right)$$

Multiplying by $e^{2\beta r_h(s,a)}$ yields

$$\sigma Q^{\pi}_h(s,a) = e^{2\beta r_h(s,a)}\,\mathrm{Var}_{S'\sim p_h(\cdot|s,a)}\left(Z^{\pi}_{h+1}(S')\right) + e^{2\beta r_h(s,a)}(p_h\,\sigma V^{\pi}_{h+1})(s,a)$$

as claimed. ■

Lemma 26 *[A normalized variance bound for an exponential transform] Let $f : \mathcal{X} \to [0,R]$ and let $\beta \in \mathbb{R}$. Define*

$$Y = e^{\beta f(X)}$$

Set

$$m = e^{\min\{0,\beta R\}} = \min\{1, e^{\beta R}\}, \qquad M = e^{\max\{0,\beta R\}} = \max\{1, e^{\beta R}\}.$$

Then $Y \in [m,M]$ a.s., $\mu \in [m,M]$, and

$$\frac{\mathrm{Var}(Y)}{\mu^2} \leqslant \frac{(e^{|\beta|R}-1)^2}{4e^{|\beta|R}}$$

Proof If $\beta = 0$ then $Y \equiv 1$ and $\mathrm{Var} = 0$, so assume $\beta \neq 0$. Since $f(X) \in [0,R]$, we have $\beta f(X) \in [\min\{0,\beta R\}, \max\{0,\beta R\}]$, hence

$$m \leqslant Y = e^{\beta f(X)} \leqslant M \quad \text{a.s.}$$

and therefore $\mu = \mathbb{E}[Y] \in [m,M]$.

By the Bhatia–Davis inequality:

$$\mathrm{Var}(Y) \leqslant (M-\mu)(\mu-m)$$

Dividing by $\mu^2 > 0$ yields

$$\frac{\mathrm{Var}(Y)}{\mu^2} \leqslant \frac{(M-\mu)(\mu-m)}{\mu^2}$$

Now consider

$$g(\mu) = \frac{(M-\mu)(\mu-m)}{\mu^2}, \qquad \mu \in [m,M]$$

Rewrite

$$g(\mu) = \frac{M+m}{\mu} - \frac{Mm}{\mu^2} - 1, \qquad g'(\mu) = -\frac{M+m}{\mu^2} + \frac{2Mm}{\mu^3} = \frac{2Mm - (M+m)\mu}{\mu^3}$$

Thus $g'(\mu) = 0$ iff $\mu^\star = \frac{2Mm}{M+m} \in [m, M]$, and g increases on $[m, \mu^\star]$ and decreases on $[\mu^\star, M]$. Hence the maximum is attained at $\mu^\star$. Substituting gives

$$g(\mu^\star) = \frac{\left(M - \frac{2Mm}{M+m}\right)\left(\frac{2Mm}{M+m} - m\right)}{\left(\frac{2Mm}{M+m}\right)^2} = \frac{(M-m)^2}{4Mm}$$

so for all $\mu \in [m, M]$

$$\frac{(M-\mu)(\mu-m)}{\mu^2} \leqslant \frac{(M-m)^2}{4Mm}$$

Finally, since $M/m = e^{|\beta|R} \geqslant 1$, we have

$$\frac{(M-m)^2}{4Mm} = \frac{\left(\frac{M}{m} - 1\right)^2}{4\frac{M}{m}} = \frac{(e^{|\beta|R} - 1)^2}{4e^{|\beta|R}}$$

■

We state lemma 11 and 12 from (Menard et al., 2021) that are used for the variance transportation:

Lemma 27 *Let $p, q \in \Sigma_S$ and f is a function defined on $\mathcal{S}$ such that $0 \leqslant f(s) \leqslant b$ for all $s \in \mathcal{S}$. If $\mathrm{KL}(p, q) \leqslant \alpha$ then*

$$\mathrm{Var}_q(f) \leqslant 2\mathrm{Var}_p(f) + 4b^2\alpha \quad \textit{and}$$
$$\mathrm{Var}_p(f) \leqslant 2\mathrm{Var}_q(f) + 4b^2\alpha$$

Lemma 28 *For $p, q \in \Sigma_S$, for f, g two functions defined on $\mathcal{S}$ such that $0 \leqslant g(s), f(s) \leqslant b$ for all $s \in \mathcal{S}$, we have that*

$$\mathrm{Var}_p(f) \leqslant 2\mathrm{Var}_p(g) + 2bp|f - g| \quad \textit{and}$$
$$\mathrm{Var}_q(f) \leqslant \mathrm{Var}_p(f) + 3b^2\|p - q\|_1,$$

where we denote the absolute operator by $|f|(s) = |f(s)|$ for all $s \in \mathcal{S}$.

We state the pseudo-counts lemma 7 that allows to go from counts to their mean the pseudo-counts and lemma 8 a standard inequality from (Menard et al., 2021)

Lemma 29 *On event $\mathcal{E}^{\mathrm{cnt}}$, for any $\alpha(\cdot, \delta)$ such that $x \mapsto \beta(\delta, x)/x$ is non-increasing for $x \geqslant 1$, $x \mapsto \beta(x, \delta)$ is non-decreasing $\forall h \in \{1, ..., H\}, (s, a) \in \mathcal{S} \times \mathcal{A}$.*

$$\forall t \in \mathbb{N}^*, \quad \frac{\alpha(n_h^t(s,a), \delta)}{n_h^t(s,a)} \wedge 1 \leqslant 4\frac{\alpha(\bar{n}_h^t(s,a)), \delta}{\bar{n}_h^t(s,a) \vee 1}$$

Lemma 30 *For $T \in \mathbb{N}^*$ and $(u_t)_{t\in\mathbb{N}^*}$ for a sequence where $u_t \in [0,1]$ and $U_t \triangleq \sum_{i=1}^{t} u_i$, we get*

$$\sum_{t=0}^{T} \frac{u_{t+1}}{U_t \vee 1} \leqslant 4\log(U_{T+1}+1).$$

Finally we state lemma 13 from (Menard et al., 2021)

Lemma 31 *Let A, B, C, D, E, and α be positive scalars such that $1 \leqslant B \leqslant E$ and $\alpha \geqslant e$. If $\tau \geqslant 0$ satisfies*

$$\tau \leqslant C\sqrt{\tau\left(A\log(\alpha\tau) + B\log(\alpha\tau)^2\right)} + D\left(A\log(\alpha\tau) + E\log(\alpha\tau)^2\right) \tag{25}$$

then

$$\tau \leqslant C^2(A+B)C_1^2 + \left(D + 2\sqrt{D}C\right)(A+E)C_1^2 + 1$$

where

$$C_1 = \frac{8}{5}\log\left(11\alpha^2(A+E)(C+D)\right)$$