Optimistic Posterior Sampling for Reinforcement Learning with Few Samples and Tight Guarantees

Daniil Tiapkin
HSE University
dtyapkin@hse.ru

Denis Belomestny
Duisburg-Essen University, HSE University
denis.belomestny@uni-due.de

Daniele Calandriello
DeepMind
dcalandriello@deepmind.com

Éric Moulines
École Polytechnique
eric.moulines@polytechnique.edu

Remi Munos
DeepMind
munos@deepmind.com

Alexey Naumov
HSE University
anaumov@hse.ru

Mark Rowland
DeepMind
markrowland@deepmind.com

Michal Valko
DeepMind
valkom@deepmind.com

Pierre Ménard
ENS Lyon
pierre.menard@ens-lyon.fr

Abstract

We consider reinforcement learning in an environment modeled by an episodic, finite, stage-dependent Markov decision process of horizon H with S states, and A actions. The performance of an agent is measured by the regret after interacting with the environment for T episodes. We propose an optimistic posterior sampling algorithm for reinforcement learning (OPSRL), a simple variant of posterior sampling that only needs a number of posterior samples logarithmic in H, S, A, and T per state-action pair. For OPSRL we guarantee a high-probability regret bound of order at most $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ ignoring poly $\log(HSAT)$ terms. The key novel technical ingredient is a new sharp anti-concentration inequality for linear forms which may be of independent interest. Specifically, we extend the normal approximation-based lower bound for Beta distributions by Alfers and Dinges [1984] to Dirichlet distributions. Our bound matches the lower bound of order $\Omega(\sqrt{H^3SAT})$, thereby answering the open problems raised by Agrawal and Jia [2017b] for the episodic setting.

1 Introduction

In reinforcement learning an agent interacts with an environment, whose underlying mechanism is unknown, by sequentially taking actions, receiving rewards, and transitioning to the next state [Sutton and Barto, 1998]. With the goal of maximizing the expected sum of the collected rewards, the agent must carefully balance between *exploring* in order to gather more information about the environment and *exploiting* the current knowledge to collect the rewards. In this paper, we are interested in solving this exploration-exploitation dilemma by injecting noise into the agent's decision-making process.

We model the environment as an episodic, finite, unknown Markov decision process (MDP) of horizon H, with S states and A actions. In particular, we consider the *stage-dependent* setting where the

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

rewards and the transition probability distributions can vary within an episode. After T episodes, the performance of an agent is measured through *regret* which is the difference between the cumulative reward the agent could have obtained by acting optimally and what the agent really obtained.

Jin et al. [2018] and Domingues et al. [2020] provide a problem-independent lower bound of order $\Omega(\sqrt{H^3SAT})$ for this setting; see also Azar et al. [2017] for a lower bound when the transitions are stage-independent.

One generic solution to the exploration-exploitation dilemma is the *principle of optimism in the face of uncertainty*. A simple way to implement this principle consists in building *upper confidence bound (UCB)* on the optimal Q-value function through the addition of *bonuses* to the rewards. This is done by either model-based algorithms [Azar et al., 2017, Dann et al., 2017, Zanette and Brunskill, 2019] or model-free algorithms [Jin et al., 2018, Zhang et al., 2020, Menard et al., 2021]; see also [Jaksch et al., 2010, Fruit et al., 2018, Talebi and Maillard, 2018] for the non-episodic setting. Notably, among others, both the upper confidence bound value iteration (UCBVI) of Azar et al. [2017] and the UCB-Advantage algorithm of Zhang et al. [2020] enjoys a problem-independent regret bound[1] of order[2] $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ that matches the aforementioned lower bound for T large enough and up to terms poly-logarithmic in H, S, A, T.

Another way is to implement the optimism by *injecting noise*. A typical example is the random least-square value iteration (RLSVI, Osband et al., 2016b, Russo, 2019) algorithm which at each episode computes new Q-values by noisy value iteration from an estimated model and then acts greedily with respect to them. In particular, a Gaussian noise is added to the reward before applying the Bellman operator to encourage exploration. Indeed, when the variance of the noise is carefully chosen, it allows to obtain optimistic Q-values with at least a fixed probability. Russo [2019] first proved a regret bound of order $\widetilde{\mathcal{O}}(H^2S^{3/2}\sqrt{AT})$ for RLSVI. Later, Xiong et al. [2021] obtained an optimal regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ for a modified version of RLSVI where the variance of the injected Gaussian noise is scaled by a term similar to the Bernstein bonuses used in UCBVI. Note that the RLSVI was also successfully extended beyond the tabular case to settings with function approximation, e.g. see Ishfaq et al., 2021, Zanette et al., 2020.

Recently, Pacchiano et al. [2021] analyzed a version of RLSVI where the Gaussian noise is replaced by a bootstrap sample of *the past rewards* and added pseudo rewards in the same fashion as Kveton et al. [2019]. The algorithm proposed by Pacchiano et al. [2021], comes with a regret bound of order $\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$.

By generalizing the Thompson sampling algorithm [Thompson, 1933] originally given for stochastic multi-armed bandit, Osband et al. [2013] propose a posterior sampling for reinforcement learning (PSRL). PSRL algorithm also relies on noise to drive exploration. The general idea behind it is to maintain a *surrogate Bayesian model* on the MDP, for instance, a Dirichlet posterior on the transition probability distribution if the rewards are known. At each episode, a new MDP is sampled (i.e., a transition probability for each state-action pair) according to the posterior distribution of the Bayesian model. Then, the agent acts optimally in this sampled MDP. As the posterior is not well concentrated in the unexplored region of the MDP, the probability that the Q-value of the sampled MDP is optimistic in this region is high. Therefore, the agent will be incentivized to explore. Although the original Thompson sampling is well-studied in the frequentist setting [Agrawal and Goyal, 2012, Kaufmann et al., 2012, Agrawal and Goyal, 2013, Zhang, 2022] and the Bayesian setting [Thompson, 1933, Russo and Roy, 2016, Russo and Van Roy, 2014], most of the analysis of PSRL only provide Bayesian regret bounds [Osband et al., 2013, Abbasi-Yadkori and Szepesvári, 2015, Osband et al., 2016b, Ouyang et al., 2017, Osband and Van Roy, 2017], i.e., when the true MDP is effectively sampled according to the prior of the surrogate Bayesian model. Despite this lack of guarantees, PSRL demonstrates competitive empirical performance in comparison to bonus-based algorithms [Osband et al., 2013, Osband and Van Roy, 2017]. Additionally, the exploration mechanism used by PSRL (and RLSVI) was successfully extended outside the tabular setting and used in deep RL environments [Osband et al., 2016a, 2018, 2019].

[1] We translate all the bounds to the *stage-dependent* setting by multiplying the regret bounds in the stage-independent setting by $\sqrt{H}$, see Jin et al. [2018].

[2] In the $\widetilde{\mathcal{O}}(\cdot)$ notation we ignore terms poly-log in H, S, A, T.

One exception to the above is the work of Agrawal and Jia [2017b] that studies PSRL from a *frequentist* perspective in the infinite-horizon, non-episodic average reward setting. In particular, they provide a regret bound[3] of order[4] $\widetilde{\mathcal{O}}(H^2 S\sqrt{AT})$ for an optimistic version of PSRL that we call SOS-OPS-RL since it switches between two types of sampling of the transitions: (1) *simple optimistic sampling*, when the number of observed transitions at a given state-action pair is too small. In this case, the sampled transition is a random mixture between the uniform distribution over the states and an empirical estimate of the true transition biased by some bonus-like terms; or if the number of observed transitions at a given state-action pair is large enough (2) *optimistic posterior sampling,* where $\widetilde{\mathcal{O}}(S)$ samples from an inflated Dirichlet posterior are used instead of one sample used in PSRL. Then, from these $\widetilde{\mathcal{O}}(S)$ sampled transition probabilities we select the most optimistic one i.e., the one leading to the largest optimal Q-value.

The key idea underpinning the analysis of SOS-OPS-RL, and PSRL-like algorithms in general, is to control the deviations of the Dirichlet posterior on the transition probability distributions. In particular, we need to show that the *posterior spreads enough to ensure optimism*. To this end, Agrawal and Jia [2017b] derive an anti-concentration bound for any fixed projection of a Dirichlet random vector. The latter result in turn relies upon an equivalent representation of a Dirichlet vector in terms of independent Beta random variables and an anti-concentration bound for the corresponding Beta distribution. However, this anti-concentration inequality is not uniformly tight, in particular its polynomial dependence on the number of states S is suboptimal.

Agrawal and Jia [2017b] conclude with two open problems. The first question is whether one can reduce the number of posterior samples required per state-action pair from $\widetilde{\mathcal{O}}(S)$ to constant or logarithmic in S. The second asks if it is possible to obtain a near-optimal regret bound and in particular to improve the dependence on S. In this paper, we *answer both of them in the affirmative* in the episodic setting. Indeed, we propose optimistic posterior sampling algorithm for reinforcement learning (OPSRL) that only requires $\widetilde{\mathcal{O}}(1)$ samples from an inflated posterior while enjoying a near-optimal problem independent regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3 SAT})$. OPSRL is a simple optimistic variant of PSRL which, in particular, does not rely at all on "simple" (bonus-based) optimistic sampling.

The essential ingredient for OPSRL's analysis is our *novel anti-concentration bound for the projections of a Dirichlet random vector* (Theorem 3.3). We base it on a tight Gaussian approximation for linear forms of a Dirichlet random vector. This latter approximation can be seen as a substantial generalization to Dirichlet distributions of the result obtained by Alfers and Dinges [1984] for the case of Beta distributions. We obtain this approximation through a refined non-asymptotic analysis of the integral representation for the density of a linear form of a Dirichlet random vector, which was first derived[5] by Tiapkin et al. [2022]. We believe that the new anti-concentration inequality presented in this work could be of independent interest, e.g., to tighten or simplify analysis of non-parametric Thompson sampling like algorithms [Riou and Honda, 2020, Baudry et al., 2021a,b] for stochastic multi-armed bandits.

- We propose the OPSRL algorithm for tabular, stage-dependent, episodic RL. It is a simple optimistic variant of the PSRL algorithm that only needs $\widetilde{\mathcal{O}}(1)$ posterior samples per state-action pair. For OPSRL, we provide a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3 SAT})$ matching the problem independent lower bound up to poly-log terms. In particular we answer positively to two open questions by Agrawal and Jia [2017b] in the episodic setting.
- We derive a new anti-concentration inequality for a linear form of a Dirichlet random vector (Theorem 3.3) which is essential for the analysis of OPSRL. This result is a generalization to the Dirichlet case of the one provided by Alfers and Dinges [1984] for Beta distributions.

[3] As acknowledged by the authors, there was a mistake in the initial submission of their work where the previously announced bound was claimed to be $\sqrt{S}$ better, see Agrawal and Jia [2017a], Qian et al. [2020]

[4] We translate all the bounds from the infinite-horizon, non-episodic average reward setting to our setting by identifying the diameter with the horizon H and multiplying the bound by $\sqrt{H}$ because of our stage-dependent transitions assumption.

[5] Note that the anti-concentration inequality proved by Tiapkin et al. [2022] based on the same integral representation is insufficient for our needs, see Remark 3.4 for a discussion.

2 Setting

We consider a finite episodic MDP $\left(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\right)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, H is the number of steps in one episode, $p_h(s'|s,a)$ is the probability transition from state s to state s' by taking the action a at step h, and $r_h(s,a) \in [0,1]$ is the bounded deterministic[6] reward received after taking the action a in state s at step h. Note that we consider the general case of rewards and transition functions that are possibly non-stationary, i.e., that are allowed to depend on the decision step h in the episode. We denote by S and A the number of states and actions, respectively.

Policy & value functions A *deterministic* policy π is a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ for all $h \in [H]$, where every π_h maps each state to a *single* action. The value functions of π, denoted by V_h^π, as well as the optimal value functions, denoted by $V_h^\star$ are given by the Bellman and the optimal Bellman equations,

$$\begin{aligned} Q_h^\pi(s,a) &= r_h(s,a) + p_h V_{h+1}^\pi(s,a) & V_h^\pi(s) &= \pi_h Q_h^\pi(s) \\ Q_h^\star(s,a) &= r_h(s,a) + p_h V_{h+1}^\star(s,a) & V_h^\star(s) &= \max_a Q_h^\star(s,a), \end{aligned}$$

where by definition, $V_{H+1}^\star \triangleq V_{H+1}^\pi \triangleq 0$. Furthermore, $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h and $\pi_h g(s) \triangleq g(s, \pi_h(s))$ denotes the composition with the policy π at step h.

Learning problem The agent, to which the transitions are *unknown* (the rewards are assumed to be known for simplicity), interacts with the environment during T episodes of length H, with a *fixed* initial state s_1.[7] Before each episode t the agent selects a policy π^t based only on the past observed transitions up to episode $t-1$. At each step $h \in [H]$ in episode t, the agent observes a state $s_h^t \in \mathcal{S}$, takes an action $\pi_h^t(s_h^t) = a_h^t \in \mathcal{A}$ and makes a transition to a new state s_{h+1}^t according to the probability distribution $p_h(s_h^t, a_h^t)$ and receives a deterministic reward $r_h(s_h^t, a_h^t)$.

Regret The quality of an agent is measured through its regret, that is the difference between what it could obtain (in expectation) by acting optimally and what it really gets,

$$\mathfrak{R}^T \triangleq \sum_{t=1}^{T} V_1^\star(s_1) - V_1^{\pi^t}(s_1)\,.$$

Counts The number of times the state action-pair (s,a) was visited in step h in the first t episodes is denoted as $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\left\{(s_h^i, a_h^i) = (s,a)\right\}$. Next, we define $n_h^t(s'|s,a) \triangleq \sum_{i=1}^t \mathbb{1}\left\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\right\}$ the number of transitions from s to s' at step h.

Improper Dirichlet distribution For $m \in \mathbb{N}^*$, the probability simplex of dimension m is denoted by Δ_m. For $\alpha \in (\mathbb{R}_{++})^{m+1}$, we denote by $\mathcal{D}\mathrm{ir}(\alpha)$ the Dirichlet distribution on Δ_m with parameter α. We also extend this distribution to improper parameter $\alpha \in (\mathbb{R}_+)^{m+1}$ such that $\sum_{i=0}^m \alpha_i > 0$ by injecting $\mathcal{D}\mathrm{ir}((\alpha_i)_{i:\alpha_i>0})$ into Δ_m. Precisely, we say that $p \sim \mathcal{D}\mathrm{ir}(\alpha)$ if $(p_i)_{i:\alpha_i>0} \sim \mathcal{D}\mathrm{ir}((\alpha_i)_{i:\alpha_i>0})$ and all other coordinates are zero.

Additional notation For $N \in \mathbb{N}_{++}$, we define the set $[N] \triangleq \{1, \ldots, N\}$. We denote the uniform distribution over this set by $\mathcal{U}\mathrm{nif}[N]$. The vector of dimension N with all entries one is $\mathbf{1}^N \triangleq (1, \ldots, 1)^\intercal$. The empirical probability distribution $\widehat{p}_h^t(s,a)$ is defined as $\widehat{p}_h^t(s'|s,a) = n_h^t(s'|s,a)/n_h^t(s,a)$ if $n_h^t(s,a) > 0$ and $\widehat{p}_h^0(s'|s,a) = 1/S$ otherwise. Appendix A references all the notation used.

[6] We study deterministic rewards to simplify the proofs but our result extend to bounded random rewards as well.

[7] As explained by Fiechter [1994] and Kaufmann et al. [2020], if the first state is sampled randomly as $s_1 \sim p$, we can simply add an artificial first state $s_{1'}$ such that for any action a, the transition probability is defined as the distribution $p_{1'}(s_{1'}, a) \triangleq p$.

3 Algorithm

In this section we describe the OPSRL algorithm. In spirit, OPSRL proceeds similarly as PSRL except that it uses several posterior samples instead and acts optimistically with respect to them, explaining the name *Optimistic Posterior Sampling for Reinforcement Learning* (OPSRL).

Optimistic pseudo-state In order to define the prior used by OPSRL, we extend the state space $\mathcal{S}$ by an absorbing pseudo-state s_0 with reward $r_h(s_0, a) \triangleq r_0 > 1$ for all h, a and transition probability distribution $p_h(s'|s_0, a) \triangleq \mathbb{1}\{s' = s_0\}$. A similar pseudo-state was already introduce in previous works, see for example Brafman and Tennenholtz [2002], Szita and Lőrincz [2008]. We denote by $\mathcal{S}' = \mathcal{S} \cup \{s_0\}$ the augmented states space and by $\Delta_{\mathcal{S}'}$ the set of probability distributions over $\mathcal{S}'$.

Pseudo-counts We define the pseudo-counts, $\overline{n}_h^t(s,a) \triangleq n_h^t(s,a) + n_0$, as the counts shifted by an initial value n_0. This shift corresponds to prior transitions to the pseudo-state, that is $\overline{n}_h^t(s'|s,a) \triangleq n_h^t(s'|s,a) + n_0 \mathbb{1}\{s' = s_0\}$. Similar to the empirical transitions, we define a pseudo-empirical transition probability distribution as $\overline{p}_h^t(s,a) = \overline{n}_h^t(s'|s,a)/\overline{n}_h^t(s,a)$.

Inflated Bayesian model Like PSRL, we define a Bayesian model on the transition probability distributions, except that the prior/posterior is inflated. The practice of inflating the posterior is common in the analysis of Thompson sampling like algorithm, see Agrawal and Jia [2017b], Abeille and Lazaric [2017]. Precisely, the inflated prior is a Dirichlet distribution $\mathcal{D}\mathrm{ir}\Big(\big(\overline{n}_h^0(s'|s,a)/\kappa\big)_{s'\in\mathcal{S}'}\Big)$ parameterized by the initial pseudo-counts, and some constant $\kappa > 0$ controlling the inflation. Thus the prior is a Dirac distribution at a deterministic transition leading to the artificial state s_0. Then the inflated posterior is also a Dirichlet distribution $\mathcal{D}\mathrm{ir}\Big(\big(\overline{n}_h^t(s'|s,a)/\kappa\big)_{s'\in\mathcal{S}'}\Big)$. Note that the prior is a proper prior (i.e., a valid probability distribution), but it will be updated in an improper way, i.e., probability transitions with no mass under the prior could get mass in the posterior, as they get positive counts.

Optimistic posterior sampling After episode t, for each state-action pair (s,a) and step $h \in [H]$ we sample J independent transition probability distributions $\widetilde{p}_h^{t,j}(s,a) \sim \mathcal{D}\mathrm{ir}\Big(\big(\overline{n}_h^t(s'|s,a)/\kappa\big)_{s'\in\mathcal{S}'}\Big)$ from the inflated posterior. Then, the Q-values are obtained by optimistic backward induction with these transitions. Precisely the value after the last step is zero $\overline{V}_{H+1}^t(s) \triangleq 0$ and the optimal Bellman equations become

$$
\begin{aligned}
\overline{Q}_h^t(s,a) &\triangleq r_h(s,a) + \max_{j\in[J]} \widetilde{p}_h^{t,j} \overline{V}_{h+1}^t(s,a)\,,\\
\overline{V}_h^t(s) &\triangleq \max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a)\,.
\end{aligned}
\tag{1}
$$

The next policy is greedy with the Q-values $\pi_h^{t+1}(s) \in \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a)$. The complete procedure of OPSRL is described in Algorithm 1 for a general family of distributions parameterized by the pseudo-counts over the transitions instead of the inflated Dirichlet prior/posterior.

3.1 Analysis

We fix $\delta \in (0,1)$ and the number of samples

$$J \triangleq \lceil c_J \cdot \log(2SAHT/\delta) \rceil,$$

where $c_J = 1/\log(2/(1+\Phi(1)))$ and $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a normal distribution. Note that J has a logarithmic dependence on S, A, H, T, and $1/\delta$.

We now state the regret bound of OPSRL with a full proof in Appendix B. and a sketch in Section 3.2.

Theorem 3.1. *Consider a parameter $\delta \in (0,1)$. Let $\kappa \triangleq 2(\log(12SAH/\delta) + 3\log(\mathrm{e}\pi(2T+1)))$, $n_0 \triangleq \lceil \kappa(c_0 + \log_{17/16}(T)) \rceil$, $r_0 \triangleq 2$, where c_0 is an absolute constant defined in (4); see Appendix B.2. Then for OPSRL, with probability at least $1-\delta$,*

$$\mathfrak{R}^T = \mathcal{O}\Big(\sqrt{H^3SATL^3} + H^3S^2AL^3\Big),$$

Algorithm 1 OPSRL

1: **Input:** Family of probability distributions $\rho : \mathbb{N}_+^{S+1} \to \Delta_{\mathcal{S}'}$ over transitions, initial pseudo-count $\overline{n}_h^0$, number of posterior samples J.
2: **for** $t \in [T]$ **do**
3: For all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, sample J independent transitions

$$\widetilde{p}_h^{t-1,j}(s,a) \sim \rho\big(\overline{n}_h^{t-1}(s'|s,a)_{s'\in\mathcal{S}'}\big), \quad j \in [J].$$

4: Optimistic backward induction: set $\overline{V}_{H+1}^{t-1}(s) = 0$ and recursively for $h \in [H]$, compute

$$\overline{Q}_h^{t-1}(s,a) = r_h(s,a) + \max_{j\in[J]}\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-1}(s,a)\}\,,$$
$$\overline{V}_h^{t-1}(s) = \max_{a\in\mathcal{A}} \overline{Q}_h^{t-1}(s,a)\,,$$
$$\pi_h^t(s) \in \arg\max_{a\in\mathcal{A}} \overline{Q}_h^{t-1}(s,a)\,.$$

5: **for** $h \in [H]$ **do**
6: Play $a_h^t = \pi_h^t(s_h^t)$.
7: Observe $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$.
8: Increment the pseudo-count $\overline{n}_h^t(s_{h+1}^t|s_h^t, a_h^t)$.
9: **end for**
10: **end for**

where $L \triangleq \mathcal{O}(\log(HSAT/\delta))$.

Computational complexity OPSRL is a model-based algorithm, and thus gets the $\mathcal{O}(HS^2A)$ space complexity as PSRL. Since we need $\widetilde{\mathcal{O}}(1)$ posterior samples per state-action pair the time complexity of OPSRL is of order $\widetilde{\mathcal{O}}(HS^2A)$ per episode, the same as PSRL up to poly-logarithmic terms. Building on the idea of Efroni et al. [2019], in Appendix F we propose the Lazy-OPSRL algorithm a more time-efficient version of OPSRL. Instead of recomputing the Q-value by backward induction before each episode, Lazy-OPSRL only performs one step of optimistic incremental planning at the visited states. It enjoys a regret bound of the same order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ as OPSRL but with an improved time-complexity per episode of $\mathcal{O}(HSA)$, see Theorem F.1 in Appendix F.

Comparison with SOS-OPS-RL and PSRL One structural difference between OPSRL and SOS-OPS-RL of Agrawal and Jia [2017a] is that OPSRL only relies on optimistic posterior sampling while SOS-OPS-RL also uses simple optimistic sampling: a mixture of the uniform distribution over the states and an empirical estimate of the true transition kernel biased by some bonus-like terms. In particular, OPSRL does not use bonus-like quantities which could lead to poor empirical performance [Osband and Van Roy, 2017]. Another important issue is the number of posterior samples. SOS-OPS-RL needs $\widetilde{\mathcal{O}}(S)$ posterior samples in order to obtain a regret bound of order $\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$ whereas OPSRL needs *only* $\widetilde{\mathcal{O}}(1)$ samples *and* obtains a better regret bound. Note that if we choose the number of posterior samples as $J = 1$ in OPSRL we recover PSRL up to two technical differences: First, the posterior is inflated in order to increase its variance. This technical trick was already used by Agrawal and Jia [2017a] and allows to guarantee optimism with a small number of posterior samples, see Section 3.2. Second, OPSRL uses a particular prior which is a Dirac distribution at a deterministic transition towards an optimistic pseudo-state. This prior is needed to control the deviations of the (inflated) posterior, see Theorem D.2.

Comparison with RLSVI Both OPSRL and RLSVI build on the same mechanism for exploration. RLSVI just adds an Gaussian noise to the Q-values whereas OPSRL injects the noise naturally via a random transition sampled from a Dirichlet distribution. As controlling the deviation of the Q-value obtained with additive Gaussian noise is not difficult, the analysis of RLSVI is relatively straightforward [Russo, 2019, Ishfaq et al., 2021]. On the contrary the analysis of OPSRL is much more involved, see Section 3.2. However, the benefit of optimistic posterior sampling in OPSRL is that it adapts *automatically* to the variance of the estimates of the transitions which is central for a regret bound with an optimal dependence on the horizon H [Azar et al., 2017]. Adapting to the variance with `RLSVI` is much more involved and artificial, see Xiong et al. [2021]. This is probably one reason why `RLSVI` performs empirically worse than `PSRL` [Osband et al., 2016a].

3.2 Proof sketch

The proof of Theorem 3.1 consists of three important steps. The first step is devoted to the approximation for tails of weighted sums of Dirichlet distribution and embodies the main technical contribution of the paper.

Step 1. Exponential and Gaussian approximation for Dirichlet distribution The first result generalizes Riou and Honda [2020] to Dirichlet distributions with real parameters. Let us first recall the definition of the minimum Kullback-Leibler divergence for $p \in \Delta_m$ where $m \in \mathbb{N}^+$, a function $f : \{0, \ldots, m\} \to [0, b]$ for some $b \in \mathbb{R}^+$ and $u \in \mathbb{R}$,

$$\mathcal{K}_{\text{inf}}(p, u, f) \triangleq \inf\{\mathrm{KL}(p, q) : q \in \Delta_m, qf \geq u\}\,,$$

where we recall that $pf \triangleq \mathbb{E}_{X\sim p} f(X)$. This quantity appears already in the analysis of non-parametric bounded multi-arm stochastic bandits, see Honda and Takemura [2010], Cappé et al. [2013]. As the Kullback-Leibler divergence, the minimum Kullback-Leibler divergence admits a variational formula by Lemma 18 of Garivier et al. [2018] up to rescaling for any $u \in (0, b)$,

$$\mathcal{K}_{\text{inf}}(p, u, f) = \max_{\lambda\in[0,1/(b-u)]} \mathbb{E}_{X\sim p}[\log(1 - \lambda(f(X) - u))]\,. \tag{2}$$

Theorem 3.2 (Exponential upper bound, see Theorem D.1)**.** *For any $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}^{m+1}_{++}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Then for any $f : \{0, \ldots, m\} \to [0, b]$ and $0 < \mu < b$, we have*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \leq \exp(-\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p}, \mu, f)).$$

The second result is devoted to a tight Gaussian lower bound for the distribution of a linear function of Dirichlet random vector. Here we follow the ideas of Alfers and Dinges [1984] and use the exact expression for the density of a linear form of Dirichlet random vector derived by Tiapkin et al. [2022].

Theorem 3.3 (Gaussian lower bound, see Theorem D.2)**.** *For any $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}^{m+1}_{++}$, define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Fix $\varepsilon \in (0, 1)$ and assume that $\alpha_0 \geq c(\varepsilon) + \log_{17/16}(\overline{\alpha})$ for $c(\varepsilon)$ defined in (11), Appendix D, and $\overline{\alpha} \geq 2\alpha_0$. Then for any $f : \{0, \ldots, m\} \to [0, b_0]$ such that $f(0) = b_0$, $f(j) \leq b < b_0/2, j \in \{1, \ldots, m\}$ and $\mu \in (\overline{p}f, b_0)$,*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \geq (1 - \varepsilon)\mathbb{P}_{g\sim\mathcal{N}(0,1)}\left[g \geq \sqrt{2\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p}, \mu, f)}\right].$$

We emphasize that increasing the parameter α_0 corresponding to the largest value of f by 1 is crucial. The same technique was used by Alfers and Dinges [1984] to derive a lower bound on the tails of the Beta distribution.

Remark 3.4. We stress that the anti-concentration inequality of Tiapkin et al. [2022, Theorem D.2] is not sufficient for our purposes; their additional factor $\overline{\alpha}^{-3/2}$ in front of the exponent makes it unusable for the analysis of `OPSRL`. Indeed, this inequality would imply $\widetilde{\mathcal{O}}(T^{3/2})$ samples from the inflated posterior in order to get optimism with high-probability, whereas with our refined bound (Theorem 3.3) we only need $\widetilde{\mathcal{O}}(1)$ posterior samples.

Proof sketch of Theorem 3.3. We start from the integral representation for the density by Tiapkin et al. [2022, Proposition D.3]. Define $Z \triangleq wf$ for $w \sim \mathcal{D}\mathrm{ir}(\alpha_0 + 1, \alpha_1, \ldots, \alpha_m)$, then for any $u \in (0, b_0)$,

$$p_Z(u) = \frac{\overline{\alpha}}{2\pi}\int_{\mathbb{R}} (1 + \mathrm{i}(b_0 - u)s)^{-1} \prod_{j=0}^{m} (1 + \mathrm{i}(f(j) - u)s)^{-\alpha_j} \mathrm{d}s.$$

One additional term $(1+\mathrm{i}(b_0-u)s)^{-1}$ comes from increasing the parameter α_0 by 1 corresponding to the value $f(0)=b_0$.

In the same spirit as it was done by Tiapkin et al. [2022], we apply the method of saddle point (see Fedoryuk, 1977, Olver, 1997) to the complex integral above. Informally, for $\alpha_0, \overline{\alpha}, b_0$ large enough the following approximation holds

$$p_Z(u) \approx \sqrt{\frac{\overline{\alpha}}{2\pi\sigma^2(1-\lambda^\star(b_0-u))^2}}\exp(-\overline{\alpha}\,\mathcal{K}_{\mathrm{inf}}(\overline{p},u,f)),$$

where $\lambda^\star$ is the unique solution to the problem (2) and $\sigma^2 = \mathbb{E}_{X\sim\overline{p}}\big[(\frac{f(X)-u}{1-\lambda^\star(f(X)-u)})^2\big]$. The formal statement can be found in Lemma D.5 of Appendix D.

Next we perform a change of variables $t^2/2 = \mathcal{K}_{\mathrm{inf}}(\overline{p},u,f)$ in the above expression to get

$$\begin{aligned}\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha_0+1,\alpha_1,\ldots,\alpha_m)}[wf\geq\mu] &\approx \int_{\mu}^{b_0}\sqrt{\frac{\overline{\alpha}}{2\pi\sigma^2(1-\lambda^\star(b_0-u))^2}}\exp(-\overline{\alpha}\,\mathcal{K}_{\mathrm{inf}}(\overline{p},u,f))\mathrm{d}u\\ &\approx \int_{\sqrt{2\,\mathcal{K}_{\mathrm{inf}}(\overline{p},\mu,f)}}^{\infty} D(u(t))\phi(t|0,\overline{\alpha})\mathrm{d}t,\end{aligned}$$

where $\phi(x|\mu,\sigma^2)$ is a density of $\mathcal{N}(\mu,\sigma^2)$ and $D(u)$ is a weight function bounded from below by 1 (see Lemma D.6 of Appendix D). This lower bound on $D(u)$ concludes the proof. □

Comparison with anti-concentration bound by Agrawal and Jia [2017b] We emphasise that our technique of deriving a Gaussian-like lower bound is substantially different from the methodology used by Agrawal and Jia [2017b]. The latter one was based on reduction of a weighted sum of Dirichlet random vector to a weighted sum of independent Beta distributed random variables and a subsequent application of the Berry-Esseen inequality, whereas our approach relies on the integral representation for the density of the corresponding linear projection of Dirichlet random vector.

In particular, the Berry-Esseen inequality is likely to be very coarse since it uses only the first three moments of the distribution and therefore generates an additional S-factor. At the same time, our analysis is much better fitted to the Dirichlet distribution and provides a very tight lower bound. The tightness of our bounds can be checked by comparing it to a similar result for the beta distribution derived in Alfers and Dinges [1984].

Step 2. Optimism Next, we apply Theorem 3.3 to prove that the estimate of Q-function $\overline{Q}_h^t$ is optimistic with high probability for our choice of inflation parameter κ and a number of posterior samples J: $\overline{Q}_h^t(s,a)\geq Q_h^\star(s,a)$ for any $(s,a,h,t)\in\mathcal{S}\times\mathcal{A}\times[H]\times[T]$.

We show that the inequalities $\max_{j\in[J]}\{\widetilde{p}_h^{t,j}V_{h+1}^\star(s,a)\}\geq p_h V_{h+1}^\star(s,a)$ hold for all $(s,a,h,t)\in\mathcal{S}\times\mathcal{A}\times[H]\times[T]$ with high probability. First, we notice that $\widetilde{p}_h^{t,j}(s,a)\sim\mathcal{D}\mathrm{ir}(\alpha_0+1,\alpha_1,\ldots,\alpha_S)$ for $\alpha_0=n_0/\kappa-1$, $\alpha_i=n_h^t(s_i|s,a)/\kappa$ and $\overline{\alpha}=(\overline{n}_h^t(s,a)-\kappa)/\kappa$. Additionally, define a probability distribution $q\in\Delta_S$ such that $q(i)=\alpha_i/\overline{\alpha}$. This distribution slightly differs from $\overline{p}_h^t(s,a)$ because of an additional $+1$ in the parameters of the Dirichlet distribution. Next, we may apply Theorem 3.3 with $\varepsilon=1/2$ and a proper choice of $n_0=n_0(\varepsilon)$,

$$\mathbb{P}_{\widetilde{p}_h^{t,j}(s,a)\sim\mathcal{D}\mathrm{ir}(\alpha_0+1,\alpha_1,\ldots,\alpha_S)}\left[\widetilde{p}_h^{t,j}V_{h+1}^\star(s,a)\geq p_hV_{h+1}^\star(s,a)\right]\geq\frac{1}{2}\left(1-\Phi\left(\sqrt{\frac{2\zeta}{\kappa}}\right)\right),$$

where $\zeta\triangleq(\overline{n}_h^t-\kappa)\,\mathcal{K}_{\mathrm{inf}}(q,p_hV_{h+1}^\star(s,a),V_{h+1}^\star)$ and $\Phi(\cdot)$ is a cumulative distribution function (CDF) of a standard normal distribution. By a concentration argument we have

$$\zeta\leq n_h^t\,\mathcal{K}_{\mathrm{inf}}(\widehat{p}_h^t(s,a),p_hV_{h+1}^\star(s,a),V_{h+1}^\star)\leq\kappa/2,$$

with high probability for an appropriate choice of $\kappa=\widetilde{\mathcal{O}}(1)$. For this step of the proof the presence of the inflation parameter κ is crucial: this parameter increases the variance of $\widetilde{p}_h^{t,j}(s,a)$ to ensure that the above inequality holds with a constant probability. Next, by taking the maximum over $J=\mathcal{O}(\log(SATH/\delta))$ samples and applying union bound, we guarantee that the inequality $\max_{j\in[J]}\{\widetilde{p}_h^{t,j}V_{h+1}^\star(s,a)\}\geq p_hV_{h+1}^\star(s,a)$ holds simultaneously for all

$(s,a,h,t) \in \mathcal{S}\times\mathcal{A}\times[H]\times[T]$ with probability at least $1-\delta/2$. The formal statement and the proof could be found in Proposition B.4 of Appendix B.2.

Finally, the standard backward induction over $h\in[H]$ concludes optimism. Indeed, the base of induction $h=H+1$ is trivial. Next, by the Bellman equations for $\overline{Q}_h^t$ and $Q_h^\star$ we have

$$\overline{Q}_h^t(s,a) - Q_h^\star(s,a) = \max_{j\in[J]}\{\widetilde{p}_h^{t,j}\overline{V}_{h+1}^t(s,a)\} - p_h V_{h+1}^\star(s,a).$$

The induction hypothesis implies $\overline{V}_{h+1}^t(s') \ge \overline{Q}_{h+1}^t(s',\pi^\star(s')) \ge Q_{h+1}^\star(s',\pi^\star(s')) = V_{h+1}^\star(s')$ for any $s'\in\mathcal{S}$. Hence,

$$\overline{Q}_h^t(s,a) - Q_h^\star(s,a) \ge \max_{j\in[J]}\{\widetilde{p}_h^{t,j}V_{h+1}^\star(s,a)\} - p_h V_{h+1}^\star(s,a) \ge 0$$

with probability at least $1-\delta/2$.

Step 3. Regret bound The rest of proof directly follows Azar et al. [2017], where UCBVI algorithm with Bernstein bonuses was analyzed. By the optimism, we have

$$\mathfrak{R}^T = \sum_{t=1}^T [V_1^\star(s_1) - V_1^{\pi^t}(s_1)] \le \sum_{t=1}^T \delta_1^t,$$

where $\delta_h^t \triangleq \overline{V}_h^{t-1}(s_h^t) - V_h^{\pi^t}(s_h^t)$. The quantity δ_h^t can be decomposed as follows using the Bellman equation for V^{π^t} and $\overline{Q}_h^{t-1}$,

$$\begin{aligned}
\delta_h^t &= \overline{Q}_h^{t-1}(s_h^t,a_h^t) - Q_h^{\pi^t}(s_h^t,a_h^t) = \max_{j\in[J]}\Big\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t)\Big\} - p_h V_{h+1}^{\pi^t}(s_h^t,a_h^t) \\
&= \underbrace{\max_{j\in[J]}\Big\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t)\Big\} - \overline{p}_h^{t-1}\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t)}_{(\mathbf{A})} + \underbrace{[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t)}_{(\mathbf{B})} \\
&+ \underbrace{[\widehat{p}_h^{t-1} - p_h][\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t,a_h^t)}_{(\mathbf{C})} + \underbrace{[\widehat{p}_h^{t-1} - p_h]V_{h+1}^\star(s_h^t,a_h^t)}_{(\mathbf{D})} \\
&+ \underbrace{p_h[\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}](s_h^t,a_h^t) - [\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}](s_{h+1}^t)}_{\xi_h^t} + \delta_{h+1}^t.
\end{aligned}$$

The terms $(\mathbf{C})$, $(\mathbf{D})$, and ξ_h^t are standard in the analysis of the optimistic algorithms. The term $(\mathbf{B})$ could be upper-bounded by $\frac{r_0\cdot n_0\cdot H}{\overline{n}_h^{t-1}(s_h^t,a_h^t)}$ and turns out to be one of second-order terms. The analysis of $(\mathbf{A})$ is novel and requires application of the Bernstein inequality for Dirichlet distributions that follows from Theorem 3.2 and is spelled out in the following lemma.

Lemma 3.5 (see Lemma C.6 in Appendix C)**.** *For any $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_m)\in\mathbb{R}_{++}^{m+1}$ define $\overline{p}\in\Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0,\ldots,m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Then for any $f\colon\{0,\ldots,m\}\to[0,b]$ such that $f(0)=b$ and $\delta\in(0,1)$,*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}\left[wf \ge \overline{p}f + 2\sqrt{\frac{\mathrm{Var}_{\overline{p}}(f)\log(1/\delta)}{\overline{\alpha}}} + \frac{3b\cdot\log(1/\delta)}{\overline{\alpha}}\right] \le \delta.$$

As opposed to Lemma C.8 of Tiapkin et al. [2022], the last result applies to Dirichlet distributions with non-integer parameters as in our case (due to the presence of the inflation parameter κ). Therefore, we see that the term $(\mathbf{A})$ can be upper bounded by a quantity which has the same role as in the analysis of UCBVI. After using the Bernstein bound, the rest of the proof follows from the analysis of UCBVI with the Bernstein bonuses and Bayes-UCBVI; see Azar et al. [2017] and Tiapkin et al. [2022].

4 Experiments

In this section we provide experiment to compare OPSRL with some baselines on simple tabular environment; see details in Appendix G. In particular, we illustrate that OPSRL is competitive with the original PSRL algorithm and outperforms bonus-based algorithms such as UCBVI.

Baselines We compare OPSRL with the following baselines: UCBVI (with Hoeffding-type bonuses) and UCBVI-B (with Bernstein-type bonuses) Azar et al. [2017], PSRL Osband et al. [2013], and RLSVI Osband et al. [2016b]. See Appendix G for full details on parameters for OPSRL and baselines.

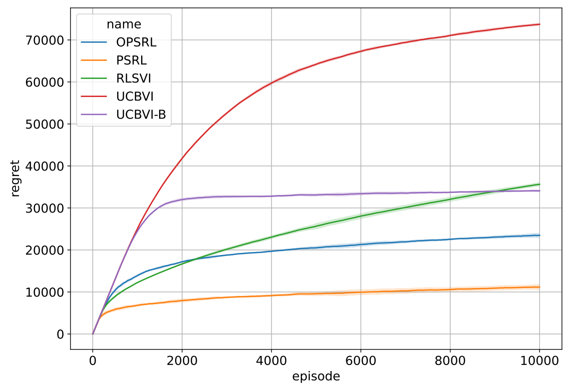

Figure 1: Regret of OPSRL and baselines on grid-world environment with 100 states and 4 action for $H = 50$ an transitions noise 0.2. We show average over 4 seeds.

Results In Figure 1, we plot the regret of the various baselines and OPSRL in the grid world environment. In this experiment, we observe that OPSRL achieves competitive results with respect to PSRL. It is not completely surprising since they share the same Bayesian model on the transitions up to the prior. We shall elaborate more on the influence of the prior in Appendix G. We also note that OPSRL outperforms UCBVI and RLSVI. This difference may be explained by the fact that OPSRL's optimism implies (in the worst case) KL bonuses as in Filippi et al. [2010]. The KL bonuses are stronger than Bernstein bonuses, see Lemma E.1, because they somehow rely on all moments of the empirical distribution rather than the first two moments as in the case of Bernstein bonuses or first moments for Hoeffding bonuses or for the variance of the Gaussian noise in RLSVI. Note also that in OPSRL, we do not have to solve the complex convex program to compute the KL bonuses Filippi et al. [2010], which could be computationally intensive.

5 Conclusion

In this work, we presented the OPSRL algorithm which can be viewed as a simple optimistic variant of the PSRL algorithm. Notably, OPSRL only needs $\widetilde{\mathcal{O}}(1)$ posterior samples per state-action. We proved that the regret of OPSRL is upper-bounded with high probability by $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$, matching the problem-independent lower-bound of order $\Omega(\sqrt{H^3SAT})$ for T large enough and up to terms poly-logarithmic in H, S, A, and T. While our work addresses the open questions raised by Agrawal and Jia [2017b] in the episodic setting, obtaining the same results in the infinite-horizon average reward setting remains an open issue. We believe that it is possible to adapt our analysis to this other setting up to some technical adjustments. Ultimately, another open question, is to obtain a high-probability regret bound for PSRL, that is, when using only a *single* posterior sample and not inflating the posterior. As a further future research direction we believe it could be interesting to obtain a model-free algorithm that relies on the same mechanism as OPSRL for exploration. Indeed, such an algorithm could avoid the use of complicated bonuses adopted by the current model-free algorithms while reducing the memory complexity of OPSRL.

Acknowledgments and Disclosure of Funding

D. Belomestny acknowledges the financial support from Deutsche Forschungsgemeinschaft (DFG), Grant Nr. 497300407. The work of D. Tiapkin, D. Belomestny and A. Naumov was prepared within the framework of the HSE University Basic Research Program. Pierre Ménard acknowledges the support of the Chaire SeqALO (ANR-20-CHIA-0020-01).

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 2–11, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108.

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/abeille17a.html.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 39.1–39.26, Edinburgh, Scotland, 25–27 Jun 2012. PMLR. URL https://proceedings.mlr.press/v23/agrawal12.html.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL https://proceedings.mlr.press/v31/agrawal13a.html.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *CoRR*, abs/1705.07041, 2017a. URL http://arxiv.org/abs/1705.07041.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017b. URL https://proceedings.neurips.cc/paper/2017/file/3621f1454cacf995530ea53652ddf8fb-Paper.pdf.

D. Alfers and H. Dinges. A normal approximation for beta and gamma tail probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(3):399–420, Feb 1984. ISSN 1432-2064. doi: 10.1007/BF00533744. URL https://doi.org/10.1007/BF00533744.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017. URL https://arxiv.org/pdf/1703.05449.pdf.

Dorian Baudry, Romain Gautron, Emilie Kaufmann, and Odalric Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 716–726. PMLR, 18–24 Jul 2021a. URL https://proceedings.mlr.press/v139/baudry21a.html.

Dorian Baudry, Patrick Saux, and Odalric-Ambrym Maillard. From optimality to robustness: Dirichlet sampling strategies in stochastic bandits. In *Neurips 2021*, 2021b.

Ronen I. Brafman and Moshe Tennenholtz. R-MAX - A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

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

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020. URL http://arxiv.org/abs/2004.05599.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 10 2021. URL https://github.com/rlberry-py/rlberry.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/25caef3a545a1fff2ff4055484f0e758-Paper.pdf.

M. V. Fedoryuk. *Metod perevala*. Izdat. "Nauka", Moscow, 1977.

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

S. Filippi, O. Cappé, and A. Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *Allerton Conference on Communication, Control, and Computing*, 2010.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

Aurélien Garivier, Hédi Hadiji, Pierre Menard, and Gilles Stoltz. Kl-ucb-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints. *arXiv preprint arXiv:1805.05071*, 2018.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT*, pages 67–79. Omnipress, 2010. ISBN 978-0-9822529-2-5. URL http://dblp.uni-trier.de/db/conf/colt/colt2010.html#HondaT10.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010. URL http://www.jmlr.org/papers/volume11/jaksch10a/jaksch10a.pdf.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Neural Information Processing Systems*, 2018. URL https://arxiv.org/pdf/1807.03765.pdf.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. *arXiv preprint arXiv:2006.05879*, jun 2020. URL http://arxiv.org/abs/2006.05879.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, ALT'12, page 199–213, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642341052. doi: 10.1007/978-3-642-34106-9_18. URL https://doi.org/10.1007/978-3-642-34106-9_18.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*, 2020. URL https://arxiv.org/pdf/2006.06294.pdf.

Branislav Kveton, Csaba Szepesvari, Sharan Vaswani, Zheng Wen, Tor Lattimore, and Mohammad Ghavamzadeh. Garbage in, reward out: Bootstrapping exploration in multi-armed bandits. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3601–3610. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kveton19a.html.

Pierre Menard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. Ucb momentum q-learning: Correcting the bias without forgetting, 2021. URL https://arxiv.org/abs/2103.01312.

Frank W. J. Olver. *Asymptotics and special functions*. AKP Classics. A K Peters, Ltd., Wellesley, MA, 1997. ISBN 1-56881-069-5. Reprint of the 1974 original [Academic Press, New York; MR0435697 (55 #8655)].

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2701–2710. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/osband17a.html.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/6a5889bb0190d0211a991f47bb19a777-Paper.pdf.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2377–2386, New York, New York, USA, 20–22 Jun 2016b. PMLR. URL https://proceedings.mlr.press/v48/osband16.html.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a7b238ba0f6502e5d6be14424b20ded-Paper.pdf.

Ian Osband, Benjamin Van Roy, Daniel J. Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. URL http://jmlr.org/papers/v20/18-339.html.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 1333–1342, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. Towards tractable optimism in model-based reinforcement learning. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1413–1423. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/pacchiano21a.html.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Concentration inequalities for multinoulli random variables. *CoRR*, abs/2001.11595, 2020. URL https://arxiv.org/abs/2001.11595.

Charles Riou and Junya Honda. Bandit algorithms based on thompson sampling for bounded reward distributions. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 777–826. PMLR, 08 Feb–11 Feb 2020. URL https://proceedings.mlr.press/v117/riou20a.html.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/451ae86722d26a608c2e174b2b2773f1-Paper.pdf.

Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of thompson sampling. *Journal of Machine Learning Research*, 17(68):1–30, 2016. URL http://jmlr.org/papers/v17/14-087.html.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/301ad0e3bd5cb1627a2044908a42fdc2-Paper.pdf.

R. Sutton and A. Barto. *Reinforcement Learning: an Introduction*. MIT press, 1998.

István Szita and András Lőrincz. The many faces of optimism: A unifying approach. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1048–1055, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390288. URL https://doi.org/10.1145/1390156.1390288.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *Algorithmic Learning Theory*, pages 770–805, 2018.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 12 1933. ISSN 0006-3444. doi: 10.1093/biomet/25.3-4.285. URL https://doi.org/10.1093/biomet/25.3-4.285.

Daniil Tiapkin, Denis Belomestny, Eric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, and Pierre Menard. From Dirichlet to rubin: Optimistic exploration in RL without bonuses. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21380–21431. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tiapkin22a.html.

Zhihan Xiong, Ruoqi Shen, and Simon S. Du. Randomized exploration is near-optimal for tabular MDP. *CoRR*, abs/2102.09703, 2021. URL https://arxiv.org/abs/2102.09703.

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, (ICML)*, 2019.

Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1954–1964. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/zanette20a.html.

Tong Zhang. Feel-good thompson sampling for contextual bandits and reinforcement learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022. doi: 10.1137/21M140924X. URL https://doi.org/10.1137/21M140924X.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/pdf/2004.10019.pdf.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section G of Appendix.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section G of Appendix.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

Table of Contents

A Notation

Table 1: Table of notation use throughout the paper

Notation	Meaning
$\mathcal{S}$	state space of size S
$\mathcal{A}$	action space of size A
H	length of one episode
T	number of episodes
J	number of posterior samples
$r_h(s,a)$	reward
$p_h(s'\|s,a)$	probability transition
$Q_h^\pi(s,a)$	Q-function of a given policy π at step h
$V_h^\pi(s)$	V-function of a given policy π at step h
$Q_h^\star(s,a)$	optimal Q-function at step h
$V_h^\star(s)$	optimal V-function at step h
$\mathfrak{R}^T$	regret
n_0	number of pseudo-samples
s_0	pseudo-state
r_0	pseudo-reward
κ	posterior inflation parameter
s_h^t	state that was visited at h step during t episode
a_h^t	action that was picked at h step during t episode
$n_h^t(s,a)$	number of visits of state-action $n_h^t(s,a) = \sum_{k=1}^t \mathbb{1}\{(s_h^k, a_h^k) = (s,a)\}$
$n_h^t(s'\|s,a)$	number of transition to s' from state-action $n_h^t(s'\|s,a) = \sum_{k=1}^t \mathbb{1}\{(s_h^k, a_h^k, s_{h+1}^k) = (s,a,s')\}$.
$\overline{n}_h^t(s,a)$	pseudo number of visits of state-action $\overline{n}_h^t(s,a) = n_h^t(s,a) + n_0$
$\overline{n}_h^t(s'\|s,a)$	pseudo number of transition to s' from state-action $\overline{n}_h^t(s'\|s,a) = n_h^t(s'\|s,a) + \mathbb{1}\{s' = s_0\} \cdot n_0$
$\widehat{p}_h^t(s'\|s,a)$	empirical probability transition $\widehat{p}_h^t(s'\|s,a) = n_h^t(s'\|s,a)/n_h^t(s,a)$
$\overline{p}_h^t(s'\|s,a)$	pseudo-empirical probability transition $\overline{p}_h^t(s'\|s,a) = \overline{n}_h^t(s'\|s,a)/\overline{n}_h^t(s,a)$
$\overline{Q}_h^t(s,a)$	upper approximation of the optimal Q-value
$\overline{V}_h^t(s,a)$	upper approximation of on the optimal V-value
$\mathbb{R}_+$	non-negative real numbers
$\mathbb{R}_{++}$	positive real numbers
$\mathbb{N}_{++}$	positive natural numbers
$[n]$	set $\{1,2,\ldots,n\}$
Δ_d	d-dimensional probability simplex: $\Delta_d = \{x \in \mathbb{R}_+^{d+1} : \sum_{j=0}^d x_j = 1\}$
$\mathbf{1}^N$	vector of dimension N with all entries one is $\mathbf{1}^N \triangleq (1,\ldots,1)$
$\\|x\\|_1$	ℓ_1-norm of vector $\\|x\\|_1 = \sum_{j=1}^m \|x_j\|$
$\\|x\\|_2$	ℓ_2-norm of vector $\\|x\\|_2 = \sqrt{\sum_{j=1}^m x_j^2}$
$\\|f\\|_2$	for $f\colon \mathsf{X} \to \mathbb{R}$, where $\|\mathsf{X}\| < \infty$ define $\\|f\\|_2 = \sqrt{\sum_{x\in\mathsf{X}} f^2(x)}$

Let $(\mathsf{X}, \mathcal{X})$ be a measurable space and $\mathcal{P}(\mathsf{X})$ be the set of all probability measures on this space. For $p \in \mathcal{P}(\mathsf{X})$ we denote by $\mathbb{E}_p$ the expectation w.r.t. p. For random variable $\xi : \mathsf{X} \to \mathbb{R}$ notation $\xi \sim p$ means $\mathrm{Law}(\xi) = p$. We also write $\mathbb{E}_{\xi\sim p}$ instead of $\mathbb{E}_p$. For independent (resp. i.i.d.) random variables $\xi_\ell \overset{\text{ind}}{\sim} p_\ell$ (resp. $\xi_\ell \overset{\text{i.i.d}}{\sim} p$), $\ell = 1,\ldots,d$, we will write $\mathbb{E}_{\xi_\ell \overset{\text{ind}}{\sim} p_\ell}$ (resp. $\mathbb{E}_{\xi_\ell \overset{\text{i.i.d}}{\sim} p}$), to denote expectation w.r.t. product measure on $(\mathsf{X}^d, \mathcal{X}^{\otimes d})$. For any $p, q \in \mathcal{P}(\mathsf{X})$ the Kullback-Leibler divergence $\mathrm{KL}(p,q)$ is given by

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

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

We write $f(S, A, H, T) = \mathcal{O}(g(S, A, H, T, \delta))$ if there exist $S_0, A_0, H_0, T_0, \delta_0$ and constant $C_{f,g}$ such that for any $S \geq S_0, A \geq A_0, H \geq H_0, T \geq T_0, \delta < \delta_0, f(S, A, H, T, \delta) \leq C_{f,g} \cdot g(S, A, H, T, \delta)$. We write $f(S, A, H, T, \delta) = \widetilde{\mathcal{O}}(g(S, A, H, T, \delta))$ if $C_{f,g}$ in the previous definition is poly-logarithmic in $S, A, H, T, 1/\delta$.

For $\lambda > 0$, we define $\mathcal{E}(\lambda)$ as an exponential distribution with a parameter λ. For $k, \theta > 0$ define $\Gamma(k, \theta)$ as a gamma-distribution with a shape parameter k and a rate parameter θ. For set X such that $|\mathsf{X}| < \infty$ define $\mathcal{U}\mathrm{nif}(\mathsf{X})$ as a uniform distribution over this set. In particular, $\mathcal{U}\mathrm{nif}[N]$ is a uniform distribution over a set $[N]$.

We fix a function $f : \{1, \ldots, m\} \mapsto [0, b]$ and recall the definition of the minimum Kullback-Leibler divergence for $p \in \Delta_{m-1}$ and $u \in \mathbb{R}$

$$\mathcal{K}_{\text{inf}}(p, u, f) \triangleq \inf\{\mathrm{KL}(p, q) : q \in \Delta_{m-1}, qf \geq u\}\,.$$

As the Kullback-Leibler divergence this quantity admits a variational formula by Lemma 18 of Garivier et al. [2018] up to rescaling for any $u \in (0, b)$

$$\mathcal{K}_{\text{inf}}(p, u, f) = \max_{\lambda \in [0, 1/(b-u)]} \mathbb{E}_{X \sim p}[\log(1 - \lambda(f(X) - u))]\,.$$

B Proof of regret bound for OPSRL

B.1 Concentration events

Let $\beta^\star, \beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta^{\mathrm{Var}} \colon (0,1) \times \mathbb{N} \to \mathbb{R}_+$, $\beta^{\mathcal{D}\mathrm{ir}} \colon (0,1) \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}_+$, and $\beta \colon (0,1) \to \mathbb{R}_+$ be some function defined later in Lemma B.1. We define the following favorable events,

$$
\begin{aligned}
\mathcal{E}^\star(\delta) &\triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \\
&\qquad \mathcal{K}_{\inf}(\widehat{p}^{\,t}_h(s,a), p_h V^\star_{h+1}(s,a), V^\star_{h+1}) \leq \frac{\beta^\star(\delta, n^t_h(s,a))}{n^t_h(s,a)} \Bigg\}, \\
\mathcal{E}^{\mathrm{KL}}(\delta) &\triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \\
&\qquad \mathrm{KL}(\widehat{p}^{\,t}_h(s,a), p_h(s,a)) \leq \frac{S \cdot \beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)} \Bigg\}, \\
\mathcal{E}^{\mathrm{conc}}(\delta) &\triangleq \Bigg\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \\
&\qquad |(\widehat{p}^{\,t}_h - p_h) V^\star_{h+1}(s,a)| \leq \sqrt{2\mathrm{Var}_{p_h}(V^\star_{h+1})(s,a) \frac{\beta(\delta, n^t_h(s,a))}{n^t_h(s,a)}} + 3H \frac{\beta(\delta, n^t_h(s,a))}{n^t_h(s,a)} \Bigg\}, \\
\mathcal{E}^{\mathcal{D}\mathrm{ir}}(\delta) &\triangleq \Bigg\{ \forall t \in [T], \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall j \in [J] : \\
&\qquad [\widetilde{p}^{\,t,j}_h - \overline{p}^{\,t}_h] \overline{V}^t_{h+1}(s,a) \leq 2\sqrt{\mathrm{Var}_{\overline{p}^{\,t}_h}[\overline{V}^t_{h+1}](s,a) \frac{\beta^{\mathcal{D}\mathrm{ir}}(\delta, T, J) \cdot \kappa}{\overline{n}^t_h(s,a)}} + 3 r_0 H \frac{\beta^{\mathcal{D}\mathrm{ir}}(\delta, T, J) \cdot \kappa}{\overline{n}^t_h(s,a)} \Bigg\}, \\
\mathcal{E}^{\mathrm{Var}}(\delta) &\triangleq \Bigg\{ \forall t \in \mathbb{N} : \quad \sum_{\ell=1}^{t} \sum_{h=1}^{H} \mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}(s^\ell_h, a^\ell_h)] \leq H^2 t + \sqrt{2H^5 t \beta^{\mathrm{Var}}(\delta, t)} + 3H^3 \beta^{\mathrm{Var}}(\delta, t) \Bigg\}, \\
\mathcal{E}(\delta) &\triangleq \Bigg\{ \sum_{t=1}^{T} \sum_{h=1}^{H} \Big| p_h [\overline{V}^{t-1}_{h+1} - V^{\pi^t}_{h+1}](s^t_h, a^t_h) - [\overline{V}^{t-1}_{h+1} - V^{\pi^t}_{h+1}](s^t_{h+1}) \Big| \leq 2 r_0 H \sqrt{2HT\beta(\delta)}, \\
&\qquad \sum_{t=1}^{T} \sum_{h=1}^{H} (1 - 1/H)^{H-h+1} \Big| p_h [\overline{V}^{t-1}_{h+1} - V^{\pi^t}_{h+1}](s^t_h, a^t_h) \\
&\qquad\qquad - [\overline{V}^{t-1}_{h+1} - V^{\pi^t}_{h+1}](s^t_{h+1}) \Big| \leq 2\mathrm{e} r_0 H \sqrt{2HT\beta(\delta)}, \Bigg\}.
\end{aligned}
$$

We also introduce the intersection of these events, $\mathcal{G}^{\mathrm{conc}}(\delta) \triangleq \mathcal{E}^\star(\delta) \cap \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}^{\mathrm{conc}}(\delta) \cap \mathcal{E}^{\mathcal{D}\mathrm{ir}}(\delta) \cap \mathcal{E}^{\mathrm{Var}}(\delta) \cap \mathcal{E}(\delta)$. We prove that for the right choice of the functions $\beta^\star, \beta^{\mathrm{KL}}, \beta^{\mathrm{conc}}, \beta, \beta^{\mathrm{Var}}$, the above events hold with high probability.

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

$$
\begin{aligned}
\beta^\star(\delta, n) &\triangleq \log(12SAH/\delta) + 3\log(\mathrm{e}\pi(2n+1)), \\
\beta^{\mathrm{KL}}(\delta, n) &\triangleq \log(12SAH/\delta) + \log(\mathrm{e}(1+n)), \\
\beta^{\mathrm{conc}}(\delta, n) &\triangleq \log(12SAH/\delta) + \log(4\mathrm{e}(2n+1)), \\
\beta^{\mathcal{D}\mathrm{ir}}(\delta, t, J) &\triangleq \log(12SAHt/\delta) + \log(J), \\
\beta^{\mathrm{Var}}(\delta, t) &\triangleq \log(48\mathrm{e}(2t+1)/\delta), \\
\beta(\delta) &\triangleq \log(48/\delta),
\end{aligned}
$$

it holds that

$$\mathbb{P}[\mathcal{E}^\star(\delta)] \geq 1-\delta/12, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1-\delta/12, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geq 1-\delta/12,$$
$$\mathbb{P}[\mathcal{E}^{\mathcal{D}\mathrm{ir}}(\delta)] \geq 1-\delta/12, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \geq 1-\delta/12, \qquad \mathbb{P}[\mathcal{E}(\delta)] \geq 1-\delta/12.$$

In particular, $\mathbb{P}[\mathcal{G}^{\mathrm{conc}}(\delta)] \geq 1-\delta/2$.

Remark B.2. Since we take $J \triangleq \Theta(\log(SAHT/\delta))$, all functions β are logarithmic in S, A, H, T, δ.

Proof. $\mathbb{P}[\mathcal{E}^\star(\delta)] \geq 1-\delta/12$ follows from Theorem C.4. Applying Theorem C.1 and the union bound over $h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \geq 1-\delta/12$. Next, by Lemma C.6 and the union bound over $h\in[H], t\in[T], (s,a)\in\mathcal{S}\times\mathcal{A}, j\in[J]$ we conclude $\mathbb{P}[\mathcal{E}^{\mathcal{D}\mathrm{ir}}(\delta)] \geq 1-\delta/12$. Theorem C.5 and the union bound over $h\in[H], (s,a)\in\mathcal{S}\times\mathcal{A}$ yield $\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \geq 1-\delta/12$. By Lemma B.2 by Tiapkin et al. [2022] we have $\mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \geq 1-\delta/12$.

To estimate $\mathbb{P}[\mathcal{E}(\delta)]$ one may apply Azuma-Hoeffding inequality. Define the following sequences for all $t\in[T], h\in[H]$

$$\bar{Z}_{t,h} \triangleq \overline{V}^{t-1}_{h+1}(s^t_{h+1}) - V^\star_{h+1}(s^t_{h+1}) - p_h[\overline{V}^{t-1}_{h+1} - V^\star_{t+1}](s^t_h, a^t_h),$$
$$\tilde{Z}_{t,h} \triangleq (1-1/H)^{H-h+1}\Big(\overline{V}^{t-1}_{h+1}(s^t_{h+1}) - V^\star_{h+1}(s^t_{h+1}) - p_h[\overline{V}^{t-1}_{h+1} - V^\star_{h+1}](s^t_h, a^t_h)\Big).$$

It is easy to see that these sequences form a martingale-difference w.r.t filtration $\mathcal{F}_{t,h} = \sigma\big\{\{(s^\ell_{h'}, a^\ell_{h'}), \ell<t, h'\in[H]\} \cup \{(s^t_{h'}, a^t_{h'}), h'\leq h\}\big\}$. Moreover, $|\bar{Z}_{t,h}| \leq 2r_0 H, |\tilde{Z}_{t,h}| \leq 2\mathrm{e}r_0 H$ for all $t\in[T]$ and $h\in[H]$. Hence, the Azuma-Hoeffding inequality implies

$$\mathbb{P}\left(\left|\sum_{t=1}^T\sum_{h=1}^H \bar{Z}_{t,h}\right| > 2r_0 H\sqrt{2tH\cdot\beta(\delta)}\right) \leq 2\exp(-\beta(\delta)) = \delta/24,$$
$$\mathbb{P}\left(\left|\sum_{t=1}^T\sum_{h=1}^H \bar{Z}_{t,h}\right| > 2\mathrm{e}r_0 H\sqrt{2tH\cdot\beta(\delta)}\right) \leq 2\exp(-\beta(\delta)) = \delta/24.$$

By the union bound, $\mathbb{P}[\mathcal{E}(\delta)] \geq 1-\delta/12$. □

Next we reproduce proof of important corollary of Lemma B.1.

Lemma B.3. *Assume conditions of Lemma B.1. Then on the event $\mathcal{E}^{\mathrm{KL}}(\delta)$, for any $f\colon\mathcal{S}\to[0,r_0H]$, $t\in\mathbb{N}, h\in[H], (s,a)\in\mathcal{S}\times\mathcal{A}$,*

$$(\widehat{p}^t_h - p_h)f(s,a) \leq \frac{1}{H}p_h f(s,a) + \frac{5r_0H^2S\cdot\beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)},$$
$$\|\widehat{p}^t_h(s,a) - p_h(s,a)\|_1 \leq \sqrt{\frac{2S\cdot\beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)}}.$$

Proof. By application of Lemma E.1 and Lemma E.2

$$\begin{aligned}(\widehat{p}^t_h - p_h)f(s,a) &\leq \sqrt{2\mathrm{Var}_{\widehat{p}^t_h}[f](s,a)\cdot\mathrm{KL}(\widehat{p}^t_h, p_h)} + \frac{2Hr_0}{3}\mathrm{KL}(\widehat{p}^t_h, p_h)\\ &\leq 2\sqrt{\mathrm{Var}_{p_h}[f](s,a)\cdot\mathrm{KL}(\widehat{p}^t_h,p_h)} + \left(2\sqrt{2}+\frac{2}{3}\right)Hr_0\,\mathrm{KL}(\widehat{p}^t_h, p_h).\end{aligned}$$

Since $0\leq f(s)\leq r_0H$

$$\mathrm{Var}_{p_h}[f](s,a) \leq p_h[f^2](s,a) \leq r_0H\cdot p_hf(s,a).$$

Finally, by a simple bound $2\sqrt{ab}\leq a+b, a,b\geq 0$, we obtain the following

$$\begin{aligned}(\widehat{p}^t_h - p_h)f(s,a) &\leq \frac{1}{H}p_hf(s,a) + (H^2 + 2\sqrt{2}r_0H + 2r_0H/3)\,\mathrm{KL}(\widehat{p}^t_h, p_h)\\ &\leq \frac{1}{H}p_hf(s,a) + 5r_0H^2\,\mathrm{KL}(\widehat{p}^t_h,p_h).\end{aligned}$$

Definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$ implies the first statement. The second statement follows directly from the combination of Pinsker's inequality and definition of $\mathcal{E}^{\mathrm{KL}}(\delta)$. □

B.2 Optimism

In this section we prove that our estimate of Q-function $\overline{Q}_h^t(s,a)$ is optimistic that is the event

$$\mathcal{E}_{\rm opt} \triangleq \left\{\forall t \in [T], h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A} : \overline{Q}_h^t(s,a) \geq Q_h^\star(s,a)\right\}. \tag{3}$$

holds with high probability on the event $\mathcal{E}^\star(\delta)$.

Define constants

$$c_0 \triangleq \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49\cdot 4\sqrt{6}}{9}\right)^2 \frac{8}{\pi} + \log_{17/16}\left(\frac{20}{32}\right) + 1, \tag{4}$$

and

$$c_J \triangleq \frac{1}{\log\left(\frac{2}{1+\Phi(1)}\right)}, \tag{5}$$

where $\Phi(\cdot)$ is a cdf of a normal distribution.

Proposition B.4. *Assume that* $J = \lceil c_J \cdot \log(2SAHT/\delta)\rceil$, $\kappa = 2\beta^\star(\delta,T)$, $r_0 = 2$, *and* $n_0 = \lceil (c_0 + \log_{17/16}(T/\kappa))\cdot\kappa\rceil$. *Then on event* $\mathcal{E}^\star$ *the following event*

$$\mathcal{E}^{\rm anticonc}(\delta) \triangleq \left\{\forall t\in[T]\ \forall h \in [H]\ \forall (s,a)\in\mathcal{S}\times\mathcal{A} : \max_{j\in[J]}\left\{\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a)\right\} \geq p_h V_{h+1}^\star(s,a)\right\}$$

holds with probability at least $1-\delta/2$.

Proof. First, we notice that $\widetilde{p}_h^{\,t,j}(s,a)$ for all fixed t,j,h,s,a have a Dirichlet distribution with parameter $(\{\overline{n}_h^t(s'|s,a)/\kappa\}_{s'\in\mathcal{S}'})$ for an extended state-space $\mathcal{S}' = \{s_0\}\cup\mathcal{S}$. Therefore, we may apply Theorem D.2 with fixed $\varepsilon = 1/2$ for $f = V_{h+1}^\star$ if we have $b_0 = r_0(H-h) \geq 2(H-h) = 2b$ and

$$\frac{n_0}{\kappa} = \alpha_0 + 1 \geq c_0 + \log_{17/16}(\overline{n}_h^t(s,a)/\kappa)$$

for a constant c_0 defined in (4). Let us define $\alpha_0 = {}^{n_0}/_{\kappa} - 1$ and $\alpha_i = {}^{n_h^t(s_i|s,a)}/_{\kappa}$ for some ordering $s_i\in\mathcal{S}$. Then we have $\overline{\alpha} = {}^{\overline{n}_h^t(s,a)}/_{\kappa} - 1$ and $\widetilde{p}_h^{\,t,j} \sim \mathcal{D}\mathrm{ir}(\alpha_0+1,\alpha_1,\ldots,\alpha_S)$. Define a distribution $q\in\Delta_S : q(i) = \alpha_i/\overline{\alpha}$. Then if $\overline{\alpha}\geq 2\alpha_0$ Theorem D.2 yields for any $u \geq qV_{h+1}^\star$

$$\mathbb{P}\left(\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a) \geq u\right) \geq \frac{1}{2}\left(1 - \Phi\left(\sqrt{\frac{2(\overline{n}_h^t(s,a)-\kappa)\,\mathcal{K}_{\rm inf}\big(q,u,V_{h+1}^\star\big)}{\kappa}}\right)\right), \tag{6}$$

where Φ is a cdf of a normal distribution.

Notice that if we have $u < qV_{h+1}^\star$ then the following bound also holds

$$\mathbb{P}\left(\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a) \geq u\right) \geq \mathbb{P}\left(\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a) \geq \overline{p}_h^t V_{h+1}^\star(s,a)\right) \geq \frac{1}{2}(1-\Phi(0)). \tag{7}$$

Since for all $u \leq qV_{h+1}^\star$ we also have $\mathcal{K}_{\rm inf}\big(q,u,V_{h+1}^\star\big) = 0$, therefore (6) holds for all $u\geq 0$ and $\overline{\alpha}\geq 2\alpha_0$.

Next we need to handle the case $\overline{\alpha} < 2\alpha_0$. In this case we have $qV_{h+1}^\star > H-h$, thus for any $0\leq u\leq H-h$

$$\mathbb{P}\left(\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a) \geq u\right) \geq \mathbb{P}_{\xi\sim B(\alpha_0+1,\overline{\alpha}-\alpha_0)}(r_0(H-h)\xi \geq u) \geq \mathbb{P}_{\xi\sim B(\alpha_0+1,\overline{\alpha}-\alpha_0)}\left(\xi \geq \frac{1}{2}\right),$$

where we first apply a lower bound $V_{h+1}^\star(s)\geq 0$ for all $s\in\mathcal{S}$ and $V_{h+1}^\star(s_0) = r_0(H-h)$, and second apply a bound $u\leq H-h$. Here we may apply the result of Alfers and Dinges [1984, Theorem 1.2''] and obtain the following lower bound that is equivalent to (7)

$$\mathbb{P}\left(\widetilde{p}_h^{\,t,j} V_{h+1}^\star(s,a) \geq u\right) \geq \Phi\left(-\mathrm{sign}(\alpha_0/\overline{\alpha} - 1/2)\cdot\sqrt{2\overline{\alpha}\,\mathrm{kl}(\alpha_0/\overline{\alpha},1/2)}\right) \geq \frac{1}{2}(1-\Phi(0)),$$

where we used $\alpha_0/\overline{\alpha} > 1/2$.

Thus, we may apply equation (6) for $u = p_h V^\star_{h+1}(s,a) \le H - h$ and any $\overline{\alpha}$

$$\mathbb{P}\Big(\widetilde{p}_h^{t,j} V^\star_{h+1}(s,a) \ge p_h V^\star_{h+1}(s,a)\Big) \ge \frac{1}{2}\left(1 - \Phi\left(\sqrt{\frac{2(\overline{n}^t_h(s,a) - \kappa)\,\mathcal{K}_{\text{inf}}\big(q, p_h V^\star_{h+1}(s,a), V^\star_{h+1}\big)}{\kappa}}\right)\right).$$

By the following relation that follows from variational formula for $\mathcal{K}_{\text{inf}}$ with rescaling of λ to $[0,1]$

$$\begin{aligned}
(\overline{n}^t_h(s,a) - \kappa)\,\mathcal{K}_{\text{inf}}(q, u, V^\star_{h+1}) &= (\overline{n}^t_h(s,a) - \kappa) \max_{\lambda \in [0,1]} \mathbb{E}_{s' \sim q}\left[\log\left(1 - \lambda \frac{V^\star_{h+1}(s') - u}{r_0(H-h) - u}\right)\right] \\
&\le \max_{\lambda \in [0,1]} (n_0 - \kappa)\log(1-\lambda) + (\overline{n}^t_h(s,a) - n_0) \max_{\lambda \in [0,1]} \mathbb{E}_{s' \sim \widehat{p}^t_h(s,a)}\left[\log\left(1 - \lambda \frac{V^\star_{h+1}(s') - u}{r_0(H-h) - u}\right)\right] \\
&\le (\overline{n}^t_h(s,a) - n_0) \max_{\lambda \in [0,1]} \mathbb{E}_{s' \sim \widehat{p}^t_h(s,a)}\left[\log\left(1 - \lambda \frac{V^\star_{h+1}(s') - u}{H - h - u}\right)\right] \\
&= (\overline{n}^t_h(s,a) - n_0)\,\mathcal{K}_{\text{inf}}(\widehat{p}^t_h(s,a), u, V^\star_{h+1}) = n^t_h(s,a)\,\mathcal{K}_{\text{inf}}(\widehat{p}^t_h(s,a), u, V^\star_{h+1}).
\end{aligned}$$

Thus, on the event $\mathcal{E}^\star$

$$(\overline{n}^t_h(s,a) - \kappa)\,\mathcal{K}_{\text{inf}}\big(q, p_h V^\star_{h+1}(s,a), V^\star_{h+1}\big) \le \beta^\star(\delta, n^t_h(s,a)) \le \beta^\star(\delta, T),$$

and, as a corollary

$$\mathbb{P}\Big(\widetilde{p}_h^{t,j} V^\star_{h+1}(s,a) \ge p_h V^\star_{h+1}(s,a) \mid \mathcal{E}^\star(\delta)\Big) \ge \frac{1}{2}\left(1 - \Phi\left(\sqrt{\frac{2\beta^\star(\delta,T)}{\kappa}}\right)\right).$$

By taking $\kappa = 2\beta^\star(\delta, T)$ we have a constant probability of being optimistic

$$\mathbb{P}\Big(\widetilde{p}_h^{t,j} V^\star_{h+1}(s,a) \ge p_h V^\star_{h+1}(s,a) \mid \mathcal{E}^\star(\delta)\Big) \ge \frac{1 - \Phi(1)}{2} \triangleq \gamma.$$

Next, using a choice $J = \lceil \log(2SAHT/\delta)/\log(1/(1-\gamma)) \rceil = \lceil c_J \cdot \log(2SAHT/\delta) \rceil$

$$\mathbb{P}\left(\max_{j \in [J]}\Big\{\widetilde{p}_h^{t,j} V^\star_{h+1}(s,a)\Big\} \ge p_h V^\star_{h+1}(s,a) \mid \mathcal{E}^\star(\delta)\right) \ge 1 - (1-\gamma)^J \ge 1 - \frac{\delta}{2SAHT}.$$

By a union bound we conclude the statement. □

Next we provide a connection between $\mathcal{E}^{\text{anticonc}}(\delta)$ and $\mathcal{E}^{\text{opt}}$.

Proposition B.5. *For any $\delta \in (0,1)$ it holds $\mathcal{E}^{\text{opt}} \subseteq \mathcal{E}^{\text{anticonc}}(\delta)$.*

Proof. We proceed by a backward induction over h. Base of induction $h = H+1$ is trivial. Next by Bellman equations for $\overline{Q}^t_h$ and $Q^\star_h$

$$[\overline{Q}^t_h - Q^\star_h](s,a) = \max_{j \in [J]}\Big\{\widetilde{p}_h^{t,j} \overline{V}^t_{h+1}(s,a)\Big\} - p_h V^\star_{h+1}(s,a).$$

By induction hypothesis we have $\overline{V}^t_{h+1}(s') \ge \overline{Q}^t_{h+1}(s', \pi^\star(s')) \ge Q^\star_{h+1}(s', \pi^\star(s')) = V^\star_{h+1}(s')$, thus

$$[\overline{Q}^t_h - Q^\star_h](s,a) \ge \max_{j \in [J]}\Big\{\widetilde{p}_h^{t,j} V^\star_{h+1}(s,a)\Big\} - p_h V^\star_{h+1}(s,a).$$

By the definition of event $\mathcal{E}^{\text{anticonc}}(\delta)$ we conclude the statement. □

Proposition B.6 (Optimism)**.** *Assume that $J = \lceil c_J \cdot \log(2SAHT/\delta) \rceil$, $\kappa = 2\beta^\star(\delta, T)$, $r_0 = 2$ and $n_0 = \lceil (c_0 + \log_{17/16}(T/\kappa)) \cdot \kappa \rceil$, where c_0 is defined in (4) and c_J is defined in (5). Then $\mathbb{P}(\mathcal{E}^{\text{opt}} \mid \mathcal{E}^\star(\delta)) \ge 1 - \delta/2$.*

B.3 Proof of Theorem 3.1

First, we define an event $\mathcal{G}(\delta) = \mathcal{G}^{\mathrm{conc}}(\delta) \cap \mathcal{E}^{\mathrm{opt}}$ where $\mathcal{G}^{\mathrm{conc}}$ defined in Lemma B.1, and $\mathcal{E}^{\mathrm{opt}}$ defined in (3). This event is handle all required concentration and anti-concentration bounds for the proof of the regret bound. Lemma B.1 and Proposition B.6 yield the following

Corollary B.7. *Let conditions of Lemma B.1 and Proposition B.6 hold. Then* $\mathbb{P}(\mathcal{G}(\delta)) \geq 1 - \delta$.

Next, denote $\delta_h^t \triangleq \overline{V}_h^{t-1}(s_h^t) - V_h^{\pi^t}(s_h^t)$ and surrogate regret $\overline{\mathfrak{R}}_h^t \triangleq \sum_{t=1}^T \delta_h^t$. To simplify notations denote $N_h^t = n_h^{t-1}(s_h^t, a_h^t)$, $N_h^t(s) = n_h^{t-1}(s|s_h^t, a_h^t)$, $\overline{N}_h^t = \overline{n}_h^{t-1}(s_h^t, a_h^t)$, $\overline{N}_h^t(s) = \overline{n}_h^{t-1}(s|s_h^t, a_h^t)$. Let

$$L = \max\{n_0/\kappa, \log(TH), \beta^\star(\delta, T), \beta^{\mathrm{KL}}(\delta, T), \beta^{\mathcal{D}\mathrm{ir}}(\delta, T, J), \beta^{\mathrm{conc}}(\delta, T), \beta(\delta), \beta^{\mathrm{Var}}(\delta, T)\}. \tag{8}$$

Under conditions Lemma B.1 and Proposition B.6, $L = \mathcal{O}(\log(SATH/\delta)) = \widetilde{\mathcal{O}}(1)$, $n_0 \leq 2L^2 = \mathcal{O}(\log^2(SATH/\delta))$, and $\kappa \leq 2L$. In what follows we will follow ideas of `UCBVI` with the Bernstein bonuses, see Azar et al. [2017], and `Bayes-UCBVI`, see Tiapkin et al. [2022].

Lemma B.8. *Assume conditions of Theorem 3.1. Then it holds on the event $\mathcal{G}(\delta)$, for any $h \in [H]$,*

$$\overline{\mathfrak{R}}_h^T \leq U_h^T \triangleq A_h^T + B_h^T + C_h^T + 4\mathrm{e}H\sqrt{2HTL} + 2\mathrm{e}H^2SA,$$

where

$$A_h^T = 3\mathrm{e}L \sum_{t=1}^T \sum_{h'=h}^H \sqrt{\mathrm{Var}_{\overline{p}_{h'}^{t-1}}[\overline{V}_{h+1}^{t-1}](s_{h'}^t, a_{h'}^t) \cdot \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}},$$

$$B_h^T = \mathrm{e}\sqrt{2L} \sum_{t=1}^T \sum_{h'=h}^H \sqrt{\mathrm{Var}_{p_{h'}}[V_{h+1}^\star](s_{h'}^t, a_{h'}^t) \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}},$$

$$C_h^T = 26H^2SL^2 \sum_{t=1}^T \sum_{h'=h}^H \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t},$$

and L is defined in (8).

Proof. Since our actions are greedy with respect to $\overline{Q}_h^{t-1}$, we have $\overline{V}_h^{t-1}(s_h^t) = \overline{Q}_h^{t-1}(s_h^t, a_h^t)$. Then by Bellman equations for V^{π^t} and $\overline{Q}_h^{t-1}$

$$\begin{aligned}
\delta_h^t &= \overline{Q}_h^{t-1}(s_h^t, a_h^t) - Q_h^{\pi^t}(s_h^t, a_h^t) = \max_{j \in [J]}\left\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-1}(s_h^t, a_h^t)\right\} - p_h V_{h+1}^{\pi^t}(s_h^t, a_h^t) \\
&= \underbrace{\max_{j \in [J]}\left\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-1}(s_h^t, a_h^t)\right\} - \overline{p}_h^{t-1}\overline{V}_{h+1}^{t-1}(s_h^t, a_h^t)}_{(\mathbf{A})} + \underbrace{[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}_{h+1}^{t-1}(s_h^t, a_h^t)}_{(\mathbf{B})} \\
&+ \underbrace{[\widehat{p}_h^{t-1} - p_h][\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t, a_h^t)}_{(\mathbf{C})} + \underbrace{[\widehat{p}_h^{t-1} - p_h]V_{h+1}^\star(s_h^t, a_h^t)}_{(\mathbf{D})} \\
&+ \underbrace{p_h[\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}](s_h^t, a_h^t) - [\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}](s_{h+1}^t)}_{\xi_h^t} + \delta_{h+1}^t.
\end{aligned}$$

This decomposition is almost equivalent to the decomposition in the proof of `Bayes-UCBVI`, the main difference is $(\mathbf{A})$ and an another value of n_0. We notice that the term ξ_h^t is an exactly the term that appears in the definition of the event $\mathcal{E}(\delta) \subseteq \mathcal{G}^{\mathrm{conc}}(\delta)$ in Lemma B.1.

Let us analyze each term in this representation under assumption $N_h^t > 0$.

Term (A). To handle this term, we use the event $\mathcal{E}^{\mathcal{D}\mathrm{ir}}(\delta) \subseteq \mathcal{G}^{\mathrm{conc}}(\delta)$

$$\max_{j \in [J]} \left\{ \widetilde{p}_h^{\,t-1,j} \overline{V}_{h+1}^{t-1}(s_h^t, a_h^t) \right\} - \overline{p}_h^{\,t-1} \overline{V}_{h+1}^{t-1}(s_h^t, a_h^t) \leq 2\sqrt{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t) \frac{2L^2}{\overline{N}_h^t}} + 3r_0 H \frac{2L^2}{\overline{N}_h^t}$$

$$\leq 3L\sqrt{\frac{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)}{\overline{N}_h^t}} + \frac{12HL^2}{\overline{N}_h^t}.$$

Term (B). To bound (**B**) we use directly a definition of $\overline{p}_h^{\,t-1}$ and $\widehat{p}_h^{\,t-1}$

$$[\overline{p}_h^{\,t-1} - \widehat{p}_h^{\,t-1}] \overline{V}_{h+1}^{t-1}(s_h^t, a_h^t) \leq \frac{n_0 r_0 H}{\overline{N}_h^t} \leq \frac{4HL^2}{\overline{N}_h^t}.$$

Term (C). Note that by Corollary B.7 the event $\mathcal{E}^{\mathrm{opt}}$ holds. We see that $[\overline{V}_{h+1}^{t-1} - V_{h+1}^\star]$ is a non-negative function and therefore Lemma B.3 is applicable for $f(s') = [\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s')$

$$[\widehat{p}_h^{\,t-1} - p_h][\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t, a_h^t) \leq \frac{1}{H} p_h [\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t, a_h^t) + \frac{5 r_0 H^2 S \cdot \beta^{\mathrm{KL}}(\delta, N_h^t)}{N_h^t}$$

$$\leq \frac{1}{H}(\xi_h^t + \delta_h^t) + \frac{10 H^2 S \cdot L}{N_h^t}.$$

Term (D). The bound on this term is guaranteed by the event $\mathcal{E}^{\mathrm{conc}}(\delta) \subseteq \mathcal{G}^{\mathrm{conc}}(\delta)$

$$(\widehat{p}_h^{\,t-1} - p_h) V_{h+1}^\star(s_h^t, a_h^t) \leq \sqrt{2\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t, a_h^t) \frac{L}{N_h^t}} + \frac{3HL}{N_h^t}.$$

All bounds on $(\mathbf{A}) - (\mathbf{D})$ yield for $N_h^t > 0$

$$\begin{aligned}
\delta_h^t &\leq \left(1 + \frac{1}{H}\right)\delta_h^t + \left(1 + \frac{1}{H}\right)\xi_h^t \\
&+ 3L\sqrt{\frac{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)}{\overline{N}_h^t}} + \sqrt{2}L \cdot \sqrt{\frac{\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t, a_h^t)}{N_h^t}} \\
&+ \frac{10 H^2 S \cdot L}{N_h^t} + \frac{16 L^2 H}{\overline{N}_h^t}.
\end{aligned}$$

Additionally, there is a trivial bound $\delta_h^t \leq 2H$ that is valid for the case $N_h^t = 0$.

Define $\gamma_{h'} = (1 + 1/H)^{H - h' + 1}$ and notice that $\gamma_{h'} \leq \mathrm{e}, \overline{N}_h^t \geq N_h^t$. Summing it up in the definition of $\overline{\mathfrak{R}}_h^T$ we obtain

$$\begin{aligned}
\overline{\mathfrak{R}}_h^T &\leq \sum_{t=1}^T \sum_{h'=h}^H \gamma_{h'} \xi_{h'}^t + \sum_{t=1}^T \sum_{h'=h}^H 2\mathrm{e}H \mathbb{1}\{N_{h'}^t = 0\} \\
&+ 3\mathrm{e}L \sum_{t=1}^T \sum_{h'=h}^H \sqrt{\mathrm{Var}_{\overline{p}_{h'}^{t-1}}[\overline{V}_{h+1}^{t-1}](s_{h'}^t, a_{h'}^t) \cdot \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}} && \triangleq A_h^T \\
&+ \mathrm{e}\sqrt{2}L \sum_{t=1}^T \sum_{h'=h}^H \sqrt{\mathrm{Var}_{p_{h'}}[V_{h'+1}^\star](s_{h'}^t, a_{h'}^t) \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}} && \triangleq B_h^T \\
&+ 26 H^2 S L^2 \sum_{t=1}^T \sum_{h'=h}^H \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t} && \triangleq C_h^T.
\end{aligned}$$

The bound on the first term is this decomposition follows from the definition of the event $\mathcal{E}(\delta) \subseteq \mathcal{G}^{\mathrm{conc}}(\delta) \subseteq \mathcal{G}(\delta)$. To bound the second term we notice that the event $\mathbb{1}\{N_{h'}^t = 0\}$ could occur no more than SAH times. □

Next we provide two important technical results. First of them is a classical result that follows from the pigeonhole principle.

Lemma B.9. *For any $H, T \ge 1$,*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{1}\{n_h^{t-1}(s_h^t, a_h^t) > 0\}}{n_h^{t-1}(s_h^t, a_h^t)} \le 2HSAL,$$

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \frac{\mathbb{1}\{n_h^{t-1}(s_h^t, a_h^t) > 0\}}{\sqrt{n_h^{t-1}(s_h^t, a_h^t))}} \le 3H\sqrt{TSA}.$$

Proof. The main observation for both inequalities follows from pigeon-hole principle: term corresponding to each state-action pair (s,a) appears in the sum exactly $n_h^{t-1}(s,a)$ times with a value $1/n$ for n increasing from 1, thanks to the indicator, to $n_h^{t-1}(s,a)$. For the first sum we use a bound on harmonic numbers, for the second one the integral bound. □

Lemma B.10. *Assume that conditions of Theorem 3.1 are fulfilled. Then it holds on the event $\mathcal{G}(\delta)$,*

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)\mathbb{1}\{N_h^t > 0\} \le 2H^2T + 2H^2U_1^T + 38H^3S^2AL^3 + 32H^3S\sqrt{2ATL},$$

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t, a_h^t) \le 2H^2T + 2H^2U_1^T + 6H^3L + 8\sqrt{2H^5TL}.$$

where U_h^T is defined in Lemma B.8.

Proof. First, apply the second inequality in Lemma E.3,

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)\mathbb{1}\{N_h^t > 0\} \le \underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)\mathbb{1}\{N_h^t > 0\}}_{(\mathbf{W})}$$
$$+ \underbrace{2r_0^2H^2\sum_{t=1}^{T}\sum_{h=1}^{H} \|\overline{p}_h^{t-1}(s_h^t, a_h^t) - p_h(s_h^t, a_h^t)\|_1 \mathbb{1}\{N_h^t > 0\}}_{(\mathbf{X})}.$$

To bound the term $(\mathbf{X})$ one may use Lemma B.3. We obtain under assumption $N_h^t > 0$

$$\|\overline{p}_h^{t-1}(s_h^t, a_h^t) - p_h(s_h^t, a_h^t)\|_1 \le \|\overline{p}_h^{t-1}(s_h^t, a_h^t) - \widehat{p}_h^{t-1}(s_h^t, a_h^t)\|_1 + \|p_h(s_h^t, a_h^t) - \widehat{p}_h^{t-1}(s_h^t, a_h^t)\|_1$$
$$\le \frac{n_0}{\overline{N}_h^t} + \sum_{s\in\mathcal{S}} N_h^t(s)\left(\frac{1}{N_h^t} - \frac{1}{\overline{N}_h^t}\right) + \sqrt{\frac{2SL}{N_h^t}} \le \frac{2SL^2}{N_h^t} + \sqrt{\frac{2SL}{N_h^t}}.$$

Since $r_0 = 2$, Lemma B.9 implies

$$(\mathbf{X}) \le 2r_0^2H^2\sum_{t=1}^{T}\sum_{h=1}^{H} \|\overline{p}_h^{t-1}(s_h^t, a_h^t) - p_h(s_h^t, a_h^t)\|_1 \mathbb{1}\{N_h^t > 0\} \le 32H^3S^2AL^3 + 24H^3S\sqrt{2ATL}.$$

Next, we bound $(\mathbf{W})$ using the first inequality of Lemma E.3

$$(\mathbf{W}) \le 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[V_{h+1}^{\pi^t}](s_h^t, a_h^t)}_{(\mathbf{Y})} + 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} r_0Hp_h\left|\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi^t}\right|(s_h^t, a_h^t)}_{(\mathbf{Z})}.$$

The term $(\mathbf{Y})$ could be bounded using definition of the event $\mathcal{E}^{\mathrm{Var}}(\delta)$. It follows that

$$(\mathbf{Y}) \le H^2T + \sqrt{2H^5TL} + 3H^3L.$$

By Proposition B.6 we have $\overline{V}^{t-1}_{h+1}(s) \ge V^{\pi^t}_{h+1}(s)$ for any $s \in \mathcal{S}$. By the definition of ξ^t_h, δ^t_h, definition of event $\mathcal{E}(\delta)$ term, and Lemma B.8 $(\mathbf{Z})$ could be bounded as follows

$$(\mathbf{Z}) \le \sum_{t=1}^{T}\sum_{h=1}^{H} 2H(\xi^t_h + \delta^t_{h+1}) \le 2r_0 H^2\sqrt{2TL} + 2H\sum_{h=1}^{H}\overline{\mathfrak{R}}^T_{h+1} \le 4H^2\sqrt{2TL} + 2H^2 U^T_1.$$

Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\overline{p}^{t-1}_h}[\overline{V}^{t-1}_{h+1}](s^t_h,a^t_h) &\le (\mathbf{X}) + 2\cdot(\mathbf{Y}) + 2\cdot(\mathbf{Z})\\ &\le 2H^2T + 2H^2U^T_1 + (32+6)H^3S^2AL^3 + (24+8)H^3S\sqrt{2ATL}\\ &\le 2H^2T + 2H^2U^T_1 + 38H^3S^2AL^3 + 32H^3S\sqrt{2ATL}.\end{aligned}$$

The first inequality in Lemma E.3 gives us a bound for the second inequality

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^\star_{h+1}](s^t_h,a^t_h) \le 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^{\pi^t}_{h+1}](s^t_h,a^t_h)}_{(\mathbf{Y})} + 2\sum_{t=1}^{T}\sum_{h=1}^{H} r_0 H p_h\left|V^\star_{h+1} - V^{\pi^t}_{h+1}\right|(s^t_h,a^t_h).$$

Since $\overline{V}^{t-1}_h \ge V^\star_h$ by Proposition B.6, the second term is bounded by $(\mathbf{Z})$. Thus

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^\star_{h+1}](s^t_h,a^t_h) \le 2(\mathbf{Y}) + 2(\mathbf{Z}) \le 2H^2T + 2H^2U^T_1 + 8\sqrt{2H^5TL} + 6H^3L.$$

□

Lemma B.11. *Assume conditions of Theorem 3.1 and Lemma B.8. Then on the event $\mathcal{G}(\delta)$ it holds*

$$A^T_1 \le 6\mathrm{e}\sqrt{H^3SAT}\cdot L^{3/2} + 6\mathrm{e}\sqrt{H^3SAU^T_1}\cdot L^{3/2} + 27\mathrm{e}H^2S^{3/2}AL^3 + 30\mathrm{e}H^2SA^{3/4}T^{1/4}L^{7/4},$$
$$B^T_1 \le 4\mathrm{e}\sqrt{H^3SAT}\cdot L + 4\mathrm{e}\sqrt{H^3SAU^T_1}\cdot L + 8\mathrm{e}H^2S^{1/2}A^{1/2}L^2 + 10\mathrm{e}H^{7/4}S^{1/2}A^{1/2}T^{1/4}L^{5/4},$$
$$C^T_1 \le 52\mathrm{e}H^3S^2AL^3 = \widetilde{\mathcal{O}}(H^3S^2A).$$

Proof. For the term A^T_1 we apply the Cauchy—Schwartz inequality, Lemma B.10, Lemma B.9 and inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}, a, b \ge 0$,

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{h=1}^{H}\sqrt{\mathrm{Var}_{\overline{p}^{t-1}_h}[\overline{V}^{t-1}_{h+1}](s^t_h,a^t_h)\frac{\mathbb{1}\{N^t_h>0\}}{N^t_h}}\\ &\quad\le \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\overline{p}^{t-1}_h}[\overline{V}^{t-1}_{h+1}](s^t_h,a^t_h)\mathbb{1}\{N^t_h>0\}}\cdot\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{1}\{N^t_h>0\}}{N^t_h}}\\ &\quad\le \sqrt{2H^2T + 2H^2U^T_1 + 38H^3S^2AL^3 + 32H^3S\sqrt{2ATL}}\cdot\sqrt{2SAHL}\\ &\quad\le 2\sqrt{H^3SATL} + 2\sqrt{H^3SAU^T_1L} + 9H^2S^{3/2}AL^2 + 10H^2SA^{3/4}T^{1/4}L^{3/4}.\end{aligned}$$

Similarly, the term B^T_1 may be estimated as follows

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H}\sqrt{\mathrm{Var}_{p_h}[V^\star_{h+1}](s^t_h,a^t_h)\frac{\mathbb{1}\{N^t_h>0\}}{N^t_h}} &\le \sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^\star_{h+1}](s^t_h,a^t_h)}\cdot\sqrt{\sum_{t=1}^{T}\sum_{h=1}^{H}\frac{\mathbb{1}\{N^t_h>0\}}{N^t_h}}\\ &\le \sqrt{2H^2T + 2H^2U^T_1 + 8\sqrt{2H^5TL} + 6H^3L}\cdot\sqrt{2SAH\cdot L}\\ &\le 2\sqrt{H^3SATL} + 2\sqrt{H^3SAU^T_1L} + 4H^2L\sqrt{SA} + 5H^{7/4}T^{1/4}L^{3/4}\sqrt{SA}.\end{aligned}$$

Finally, to estimate C_1^T we apply Lemma B.9. We obtain

$$C_1^T \le 26\mathrm{e}H^2S \cdot L^2 \cdot 2SAHL \le 52\mathrm{e}H^3S^2AL^3.$$

□

Proof of Theorem 3.1. By Corollary B.7 event $\mathcal{G}(\delta)$ holds with probability at least $1-\delta$. Next we assume that this event holds. Then we have two cases: $T < H^2S^2AL^3$ and $T \ge H^2S^2AL^3$. In the first case the regret is trivially bounded by $\mathfrak{R}^T \le H^3S^2AL^3$. Thus it is sufficient to analyze only the second case.

By Proposition B.6 and Lemma B.8

$$\begin{aligned}\mathfrak{R}^T &= \sum_{t=1}^T V_h^\star(s_1^t) - V_h^{\pi^t}(s_1^t) \le \sum_{t=1}^T \overline{V}_h^{t-1}(s_1^t) - V_h^{\pi^t}(s_1^t) = \overline{\mathfrak{R}}_1^T \\ &\le U_1^T = A_1^T + B_1^T + C_1^T + 4\mathrm{e}\sqrt{2H^3TL} + 2\mathrm{e}H^2SA.\end{aligned} \tag{9}$$

Next, under our condition on T we can simplify expressions for the bounds of A_1^T and B_1^T. Indeed, $T \ge H^2S^2AL^3$ implies that

$$H^{7/4}S^{1/2}A^{1/2}L^{5/4} \cdot T^{1/4} \le H^2SA^{3/4}L^{7/4} \cdot T^{1/4} \le \sqrt{H^3SAT}L^{3/2}.$$

Furthermore,

$$H^2S^{3/2}AL^3 \le H^3S^2AL^3, \qquad H^2S^{1/2}A^{1/2}L^2 \le H^3S^2AL^3, \qquad \sqrt{2H^3TL} \le \sqrt{2H^3SAT} \cdot L.$$

We obtain the following bounds

$$\begin{aligned}A_1^T &\le 36\mathrm{e}\sqrt{H^3SAT} \cdot L^{3/2} + 6\mathrm{e}\sqrt{H^3SAU_1^T} \cdot L^{3/2} + 27\mathrm{e}H^3S^2AL^3,\\ B_1^T &\le 14\mathrm{e}\sqrt{H^3SAT} \cdot L + 4\mathrm{e}\sqrt{H^3SAU_1^T} \cdot L + 8\mathrm{e}H^3S^2AL^3,\\ C_1^T &\le 52\mathrm{e}H^3S^2AL^3.\end{aligned}$$

Hence, using bound $H^2SA \le H^3S^2A$,

$$\begin{aligned}U_1^T &\le 50\mathrm{e}\sqrt{H^3SAT} \cdot L^{3/2} + 10\mathrm{e}\sqrt{H^3SAU_1^T} \cdot L^{3/2} + 89\mathrm{e}H^3S^2AL^3 + 4\mathrm{e}\sqrt{2} \cdot \sqrt{H^3TL}\\ &\le 56\mathrm{e}\sqrt{H^3SAT} \cdot L^{3/2} + 10\mathrm{e}\sqrt{H^3SAU_1^T} \cdot L^{3/2} + 89\mathrm{e}H^3S^2AL^3.\end{aligned}$$

This is a quadratic inequality in U_1^T. Solving this inequality and using inequality $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, $a, b \ge 0$, we obtain

$$U_1^T \le 108\mathrm{e}\sqrt{H^3SATL^3} + 178\mathrm{e}H^3S^2AL^3 + 200\mathrm{e}^2H^3SAL^3.$$

The last inequality and (9) imply that

$$\mathfrak{R}^T = \mathcal{O}\Big(\sqrt{H^3SATL^3} + H^3S^2AL^3\Big).$$

□

C Deviation inequalities

C.1 Deviation inequality for categorical distributions

Next, we restate the deviation inequality for categorical distributions by Jonsson et al. [2020, Proposition 1]. Let $(X_t)_{t\in\mathbb{N}^\star}$ be i.i.d. samples from a distribution supported on $\{1,\dots,m\}$, of probabilities given by $p \in \Delta_{m-1}$, where Δ_{m-1} is the probability simplex of dimension $m-1$. We denote by $\widehat{p}_n$ the empirical vector of probabilities, i.e., for all $k \in \{1,\dots,m\}$,

$$\widehat{p}_{n,k} = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{1}\{X_\ell = k\}.$$

Note that an element $p \in \Delta_{m-1}$ can be seen as an element of $\mathbb{R}^{m-1}$ since $p_m = 1 - \sum_{k=1}^{m-1} p_k$. This will be clear from the context.

Theorem C.1. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0,1]$,*

$$\mathbb{P}(\exists n \in \mathbb{N}^\star,\, n\,\mathrm{KL}(\widehat{p}_n, p) > \log(1/\delta) + (m-1)\log(e(1+n/(m-1)))) \le \delta.$$

C.2 Deviation inequality for categorical weighted sum

We fix a function $f : \{1,\dots,m\} \to [0,b]$ and recall the definition of the minimum Kullback-Leibler divergence for $p \in \Delta_{m-1}$ and $u \in \mathbb{R}$

$$\mathcal{K}_{\text{inf}}(p,u,f) = \inf\{\mathrm{KL}(p,q) : q \in \Delta_{m-1}, qf \ge u\}\,.$$

As the Kullback-Leibler divergence this quantity admits a variational formula.

Lemma C.2 (Lemma 18 by Garivier et al., 2018)**.** *For all $p \in \Delta_{m-1}$, $u \in [0,b)$,*

$$\mathcal{K}_{\textit{inf}}(p,u,f) = \max_{\lambda\in[0,1]} \mathbb{E}_{X\sim p}\left[\log\left(1 - \lambda\frac{f(X)-u}{b-u}\right)\right],$$

moreover if we denote by $\lambda^\star$ the value at which the above maximum is reached, then

$$\mathbb{E}_{X\sim p}\left[\frac{1}{1-\lambda^\star\frac{f(X)-u}{b-u}}\right] \le 1\,.$$

Remark C.3. Contrary to Garivier et al. [2018] we allow that $u = 0$ but in this case Lemma C.2 is trivially true, indeed

$$\mathcal{K}_{\text{inf}}(p,0,f) = 0 = \max_{\lambda\in[0,1]} \mathbb{E}_{X\sim p}\left[\log\left(1-\lambda\frac{f(X)}{b}\right)\right].$$

We are now ready to restate the deviation inequality for the $\mathcal{K}_{\text{inf}}$ by Tiapkin et al. [2022] which is a self-normalized version of Proposition 13 by Garivier et al. [2018].

Theorem C.4. *For all $p \in \Delta_{m-1}$ and for all $\delta \in [0,1]$,*

$$\mathbb{P}\big(\exists n \in \mathbb{N}^\star,\, n\,\mathcal{K}_{\textit{inf}}(\widehat{p}_n, pf, f) > \log(1/\delta) + 3\log(e\pi(1+2n))\big) \le \delta.$$

C.3 Deviation inequality for bounded distributions

Below, we restate the self-normalized Bernstein-type inequality by Domingues et al. [2020]. 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 $|Y_t| \le b$ and centered $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0$. Consider the following quantities

$$S_t \triangleq \sum_{s=1}^{t} w_s Y_s, \quad V_t \triangleq \sum_{s=1}^{t} w_s^2 \cdot \mathbb{E}\big[Y_s^2|\mathcal{F}_{s-1}\big], \quad \text{and} \quad W_t \triangleq \sum_{s=1}^{t} w_s$$

and let $h(x) \triangleq (x+1)\log(x+1) - x$ be the Cramér transform of a Poisson distribution of parameter 1.

Theorem C.5 (Bernstein-type concentration inequality)**.** *For all $\delta > 0$,*

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

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

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

C.4 Deviation inequality for Dirichlet distribution

Below we provide the Bernstein-type inequality for weighted sum of Dirichlet distribution, using a generalized result on upper bound on tails for linear statistics on Dirichlet distribution (Theorem D.1).

Lemma C.6. *[Generalization of Lemma C.8 by Tiapkin et al. [2022]] For any $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_{++}^{m+1}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Then for any $f \colon \{0, \ldots, m\} \to [0, b]$ such that $f(0) = b$ and $\delta \in (0,1)$*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}\left[wf \geq \overline{p}f + 2\sqrt{\frac{\mathrm{Var}_{\overline{p}}(f)\log(1/\delta)}{\overline{\alpha}}} + \frac{3b\cdot\log(1/\delta)}{\overline{\alpha}}\right] \leq \delta.$$

Remark C.7. The only difference with the result of Lemma C.8 by Tiapkin et al. [2022] is allowing to parameters of Dirichlet distribution being non-integer.

Proof. Fix $\delta \in (0,1)$ and let $\mu \in (\overline{p}f, b)$ be such that

$$\mathcal{K}_{\inf}(\overline{p}, \mu, f) = \overline{\alpha}^{-1}\log(1/\delta).$$

Note that such μ exists by the continuity of $\mathcal{K}_{\inf}$ w.r.t. the second argument, see Honda and Takemura [2010, Theorem 7]. By Tiapkin et al. [2022, Lemma C.7] there exists q such that $\overline{p} \ll q$, $qf = \mu$ and $\mathrm{KL}(\overline{p}, q) = \overline{\alpha}^{-1}\log(1/\delta)$. By Theorem D.1

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq qf] = \mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \leq \exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p}, \mu, f)) = \delta. \tag{10}$$

By Lemma E.1

$$qf - \overline{p}f \leq \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(\overline{p}, q)}.$$

By Lemma E.2, $\mathrm{Var}_q(f) \leq 2\mathrm{Var}_{\overline{p}}(f) + 4b^2\,\mathrm{KL}(\overline{p}, q)$. The last two inequalities and (10) imply that

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}\left[wf - \overline{p}f \geq \sqrt{4\mathrm{Var}_{\overline{p}}(f)\,\mathrm{KL}(\overline{p}, q)} + 2b\sqrt{2}\cdot\mathrm{KL}(\overline{p}, q)\right] \leq \delta.$$

□

D Exponential and Gaussian approximations of Dirichlet distribution

In this section we present result on approximation of a tail probabilities for linear statistics of Dirichlet distribution from above by tails of exponential distribution and from below by tails of Gaussian distribution.

The proof of upper bound generalizes proof of Baudry et al. [2021b] to non-integer parameters using exactly the same technique; see also Riou and Honda [2020].

Theorem D.1 (Upper bound). *For any $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_{++}^{m+1}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Then for any $f\colon \{0, \ldots, m\} \to [0, b]$ and $0 < \mu < b$ and*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \leq \exp(-\overline{\alpha}\,\mathcal{K}_{inf}(\overline{p}, \mu, f)).$$

Proof. The statement is trivial for $\mu \leq \overline{p}f$ since $\mathcal{K}_{\inf}(\overline{p}, \mu, f) = 0$. Assume that $\mu > \overline{p}f$. It is well know fact that $w \sim \mathcal{D}\mathrm{ir}(\alpha)$ may be represented as follows

$$w \triangleq \left(\frac{Y_0}{V_m}, \frac{Y_1}{V_m}, \ldots, \frac{Y_m}{V_m}\right),$$

where $Y_\ell \overset{\mathrm{ind}}{\sim} \Gamma(\alpha_\ell, 1), \ell = 0, \ldots, m$ and $V_m = \sum_{\ell=0}^m Y_\ell$. Let us fix $\lambda \in [0, 1/(b-u))$ and proceed by the changing of measure argument

$$\begin{aligned}
\mathbb{P}(wf \geq \mu) &= \mathbb{P}\left(\sum_{\ell=0}^m Y_\ell f(\ell) \geq \sum_{\ell=0}^m Y_\ell \mu\right) = \mathbb{E}_{Y_\ell \overset{\mathrm{ind}}{\sim}\Gamma(\alpha_\ell,1)}\left[\mathbb{1}\left\{\sum_{\ell=0}^m Y_\ell(f(\ell)-\mu) \geq 0\right\}\right] \\
&= \mathbb{E}_{\hat{Y}_\ell \overset{\mathrm{ind}}{\sim}\Gamma(\alpha_\ell, 1-\lambda(f(\ell)-\mu))}\left[\mathbb{1}\left\{\sum_{\ell=0}^m \hat{Y}_\ell(f(\ell)-\mu) \geq 0\right\}\prod_{\ell=0}^m \frac{\exp(-\lambda\hat{Y}_\ell(f(\ell)-\mu))}{(1-\lambda(f(\ell)-\mu))^{\alpha_\ell}}\right] \\
&= \exp\left(-\sum_{\ell=0}^m \alpha_\ell \log(1-\lambda(f(\ell)-\mu))\right) \\
&\quad \cdot \mathbb{E}_{\hat{Y}_\ell \overset{\mathrm{ind}}{\sim}\Gamma(\alpha_\ell, 1-\lambda(f(\ell)-\mu))}\left[\mathbb{1}\left\{\sum_{\ell=0}^m \hat{Y}_\ell(f(\ell)-\mu) \geq 0\right\} \mathrm{e}^{-\lambda\sum_{\ell=0}^m \hat{Y}_\ell(f(\ell)-\mu)}\right] \\
&\leq \exp\left(-\overline{\alpha}\sum_{\ell=0}^m \overline{p}_\ell \log(1-\lambda(f(\ell)-\mu))\right) = \exp(-\overline{\alpha}\mathbb{E}_{X\sim\overline{p}}[\log(1-\lambda(f(X)-\mu)]).
\end{aligned}$$

Since the previous inequality is true for all $\lambda \in [0, 1/(b-\mu))$, then the variational formula (Lemma C.2) allows to conclude

$$\begin{aligned}
\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] &\leq \exp\left(-\overline{\alpha}\sup_{\lambda\in[0,1/(b-\mu))}\mathbb{E}_{X\sim\overline{p}}[\log(1-\lambda(f(X)-\mu)]\right) \\
&= \exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p}, \mu, f)).
\end{aligned}$$

□

The proof of lower bound extends the approach of Tiapkin et al. [2022] by ideas of Alfers and Dinges [1984] and gives much more exact bounds. Define

$$c_0(\varepsilon) = \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49\cdot 4\sqrt{6}}{9}\right)^2 \frac{2}{\pi\cdot\varepsilon^2} + \log_{17/16}\left(\frac{5}{32\cdot\varepsilon^2}\right). \tag{11}$$

Theorem D.2 (Lower bound). *For any $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_{++}^{m+1}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Let $\varepsilon \in (0, 1)$. Assume that $\alpha_0 \geq c_0(\varepsilon) + \log_{17/16}(\overline{\alpha})$ for $c_0(\varepsilon)$ defined in (11), and $\overline{\alpha} \geq 2\alpha_0$. Then for any $f\colon \{0, \ldots, m\} \to [0, b_0]$ such that $f(0) = b_0$, $f(j) \leq b < b_0/2, j \in \{1, \ldots, m\}$ and $\mu \in (\overline{p}f, b_0)$*

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \geq (1-\varepsilon)\mathbb{P}_{g\sim\mathcal{N}(0,1)}\left[g \geq \sqrt{2\overline{\alpha}\,\mathcal{K}_{inf}(\overline{p}, \mu, f)}\right].$$

In the further subsections we are going to prove this theorem.

D.1 Proof of Theorem D.2

First, we restate the result of Tiapkin et al. [2022] on representation of the density of linear statistic of Dirichlet distribution.

Proposition D.3 (Proposition D.3 of Tiapkin et al. [2022])**.** *Let $f \in \mathrm{F}_m(b)$ and $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_{++}^{m+1}$ such that $\overline{\alpha} = \sum_{j=0}^m \alpha_j > 1$. Let $w \sim \mathcal{D}\mathrm{ir}(\alpha)$ and assume that $Z = wf$ is not degenerate. Then for any $0 \le u < b_0$*

$$p_Z(u) = \frac{\overline{\alpha}}{2\pi} \int_{\mathbb{R}} (1 + \mathrm{i}(b_0 - u)s)^{-1} \prod_{j=0}^m (1 + \mathrm{i}(f(j) - u)s)^{-\alpha_j} \mathrm{d}s.$$

Next we proceed in the same spirit as an approach of Tiapkin et al. [2022] and apply the method of saddle point (see Fedoryuk [1977], Olver [1997]) to derive an asymptotically tight approximation. However, in our case we have to extract one additional term in from of the product.

Proposition D.4. *Let $f \in \mathrm{F}_m(b_0, b)$ and let $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ be a fixed vector with $\alpha_0 \ge 2$. Then for any $u \in (\overline{p}f, b_0)$,*

$$\int_{\mathbb{R}} \frac{\prod_{\ell=0}^m (1 + \mathrm{i}(f(\ell) - u)s)^{-\alpha_\ell}}{(1 + \mathrm{i}(b_0 - u)s)} \mathrm{d}s = \left(\sqrt{\frac{2\pi}{\overline{\alpha}\, \sigma^2}} - R_1(\alpha) + R_2(\alpha) \right) \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\mathit{inf}}(\overline{p}, u, f))}{1 - \lambda^\star(b_0 - u)} + R_3(\alpha),$$

where

$$\begin{aligned}
\sigma^2 &= \mathbb{E}_{X \sim \overline{p}} \left[\left(\frac{f(X) - u}{1 - \lambda^\star(f(X) - u)} \right)^2 \right], \\
|R_1(\alpha)| &\le \frac{c_1}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2\, c_\kappa \alpha_0\, \overline{\alpha}}}, \\
|R_2(\alpha)| &\le \frac{c_2}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2\, \overline{\alpha}\, \alpha_0}}, \\
|R_3(\alpha)| &\le c_3 \cdot \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\mathit{inf}}(\overline{p}, u, f))}{1 - \lambda^\star(b_0 - u)} \cdot \frac{1 - \lambda^\star(b_0 - u)}{b_0 - u} \exp(-c_\kappa \alpha_0)
\end{aligned}$$

with $c_1 = 2\sqrt{2}, c_2 = \left(8 + \frac{49\sqrt{6}}{9} \frac{b_0}{b_0 - \overline{p}f}\right), c_3 = \frac{\sqrt{5\pi}}{2}, c_\kappa = 1/2 \cdot \log\left(1 + \frac{1}{4}\left(\frac{b_0 - \overline{p}f}{b_0}\right)^2\right)$ and $\lambda^\star$ being a solution to the optimization problem

$$\lambda^\star(\overline{p}, u, f) = \mathop{\arg\max}_{\lambda \in [0, 1/(b_0 - u)]} \mathbb{E}_{X \sim \overline{p}}[\log(1 - \lambda(f(X) - u))].$$

Proof. We start from the rewriting the integral in the form that allows us to apply saddle point method,

$$\begin{aligned}
I &= \int_{\mathbb{R}} \frac{\prod_{j=0}^n (1 + \mathrm{i}(f(j) - u)s)^{-\alpha_j}}{1 - \mathrm{i}(b_0 - u)s} \mathrm{d}s = \int_{\mathbb{R}} \frac{\exp\left(-\overline{\alpha} \sum_{j=0}^m \overline{p}_j \log(1 + \mathrm{i}(f(j) - u)s)\right)}{1 - \mathrm{i}(b_0 - u)s} \mathrm{d}s \\
&= \int_{\mathbb{R}} (1 - \mathrm{i}(b_0 - u)s)^{-1} \exp(-\overline{\alpha}\, \mathbb{E}_{X \sim \overline{p}}[\log(1 + \mathrm{i}(f(X) - u)s)]) \,\mathrm{d}s. \qquad (12)
\end{aligned}$$

Since the analysis of the suitable integration contour depends only on the function under exponent, we may directly switch to the contour $\gamma^\star = \mathrm{i}\lambda^\star + \mathbb{R}$ as it was stated in Tiapkin et al. [2022].

Next we continue following approach of Tiapkin et al. [2022] and denote the following functions

$$\begin{aligned}
T(s) &= \mathbb{E}[\log(1 - \lambda^\star(f(X) - u) + \mathrm{i}s(f(X) - u))], \\
P(s) &= \frac{1}{1 - \lambda^\star(b_0 - u) + \mathrm{i}s(b_0 - u)},
\end{aligned}$$

a cut-off parameter $K > 0$, and define $\kappa_1 = T(-K) - T(0)$, $\kappa_2 = T(K) - T(0)$. Similarly to Chapter 4 (Section 6) by Olver [1997], we define the change of variables $v_1 = T(-s) - T(0), v_2 = T(s) - T(0)$ and the implicit functions $q_1(v_1) = \frac{P(-s)}{T'(-s)}$, $q_2(v_2) = \frac{P(s)}{T'(s)}$. Notice that these functions differs from ones defined in Tiapkin et al. [2022] due to the presence of an additional multiplier $P(s)$. Using the first order Taylor expansion, we can write $q_1(v_1) = \frac{P(0)}{\sqrt{2T''(0)\cdot v_1}} + r_1(v_1)$, $q_2(v_2) = \frac{P(0)}{\sqrt{2T''(0)\cdot v_2}} + r_2(v_2)$. Then we have the following decomposition

$$I = \int_{-\infty}^{\infty} P(s)\exp(-\overline{\alpha}\, T(s))\,\mathrm{d}s = \left(P(0)\cdot\sqrt{\frac{2\pi}{\overline{\alpha}\, T''(0)}} - R_1(\alpha) + R_2(\alpha)\right)\exp(-\overline{\alpha}\, T(0)) + R_3(\alpha),$$

where

$$\begin{aligned}
R_1(\alpha) &= \left(\Gamma\left(\frac{1}{2}, \kappa_1\, \overline{\alpha}\right) + \Gamma\left(\frac{1}{2}, \kappa_2\, \overline{\alpha}\right)\right)\frac{P(0)}{\sqrt{2T''(0)\,\overline{\alpha}}},\\
R_2(\alpha) &= \int_0^{\kappa_1} e^{-\overline{\alpha} v_1} r_1(v_1)\,\mathrm{d}v_1 + \int_0^{\kappa_2} e^{-\overline{\alpha} v_2} r_2(v_2)\,\mathrm{d}v_2,\\
R_3(\alpha) &= \int_{\mathbb{R}\setminus[-K,K]} P(s)\exp(-\overline{\alpha}\, T(s))\,\mathrm{d}s,
\end{aligned}$$

where $\Gamma(\alpha, x)$ is an upper incomplete gamma function and integration w.r.t. v_1, v_2 is performed over the straight lines connecting the points 0 and κ_1, κ_2, respectively. Define $\sigma^2 = T''(0)$.

Term R_2. We will start from upper bounding on remainder terms in Taylor-like expansions $r_2(v)$

$$\begin{aligned}
|r_2(v)| &= \left|\frac{P(s)}{T'(s)} - \frac{P(0)}{\sqrt{2T''(0)(T(s)-T(0))}}\right|\\
&\le P(0)\left|\frac{1}{T'(s)} - \frac{1}{\sqrt{2T''(0)(T(s)-T(0))}}\right| + \frac{|P(s)-P(0)|}{|T'(s)|}\\
&= P(0)\cdot \bar{r}_2(v) + \tilde{r}_2(v).
\end{aligned}$$

Analysis of the term $\bar{r}_2(v)$ was performed in Tiapkin et al. [2022] under the choice $1/(2K) = \max\left\{\frac{b_0-u}{1-\lambda^\star(b_0-u)}, \frac{u}{1+\lambda^\star u}\right\}$ and the upper bound $\kappa = \operatorname{Re}\kappa_2 = \operatorname{Re}\kappa_1 \ge c_\kappa \cdot \frac{\alpha_0}{\overline{\alpha}}$ with $c_\kappa = 1/2 \cdot \log\left(1 + \frac{1}{4}\left(\frac{b_0 - \overline{p}f}{b_0}\right)^2\right)$ led to

$$\bar{r}_2(v) \le \frac{49\sqrt{6}}{36\sqrt{\sigma^2}} \cdot \sqrt{\frac{\overline{\alpha}}{\alpha_0}\frac{b_0}{b_0 - \overline{p}f}}.$$

Our next goal is to analyze the second term $\tilde{r}_2(v)$. We apply Taylor expansions of the form $T'(s) = T''(0)s + T'''(\xi_2)s^2/2$ and $P(s) = P(0) + P'(\eta)s$ to derive

$$\tilde{r}_2(v) = \left|\frac{P(s)-P(0)}{T'(s)}\right| = \frac{|P'(\eta)\cdot s|}{|T''(0)s + T'''(\xi_2)s^2/2|} \le \frac{\sup_{\eta\in(0,s)}|P'(\eta)|}{|T''(0) + T'''(\xi_2)s/2|}.$$

First note that $P'(\eta)$ maximizes at $\eta = 0$, since

$$P'(\eta) = \frac{b_0 - u}{(1 - \lambda^\star(b_0-u) + \mathrm{i}\eta(b_0-u))^2}.$$

Next by defining a random variable $Y_s = \frac{f(X)-u}{1-\lambda^\star(f(X)-u)+is(f(X)-u)}$ and due to our choice of K we conclude that

$$|T''(0) + T'''(\xi_2)s/2| \ge \mathbb{E}[Y_0^2] - s\mathbb{E}[|Y_0|^3] \ge \mathbb{E}[Y_0^2]/2 = \sigma^2/2.$$

It yields

$$\tilde{r}_2(v) \le \frac{2(b_0-u)}{(1-\lambda^\star(b_0-u))^2\sigma^2} = \frac{2}{(1-\lambda^\star(b_0-u))\sqrt{\sigma^2}}\sqrt{\frac{\frac{(b_0-u)^2}{(1-\lambda^\star(b_0-u))^2}}{\mathbb{E}[Y_0^2]}}.$$

By a bound

$$\mathbb{E}[Y_0^2] = \sum_{i=0}^{m} \frac{\alpha_i}{\overline{\alpha}} \cdot \left(\frac{f(i) - u}{1 - \lambda^\star(f(i) - u)}\right)^2 \geq \frac{\alpha_0}{\overline{\alpha}} \frac{(b_0 - u)^2}{(1 - \lambda^\star(b_0 - u))^2}$$

we obtain

$$\tilde{r}_2(v) \leq \frac{2}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2}} \sqrt{\frac{\overline{\alpha}}{\alpha_0}}$$

and

$$|r_2(v)| \leq \frac{1}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2}} \sqrt{\frac{\overline{\alpha}}{\alpha_0}} \left(2 + \frac{49\sqrt{6}}{36} \frac{b_0}{b_0 - \overline{p}f}\right).$$

A similar bound also holds for $r_1(v)$ by symmetry. Finally, due to bound on κ and $\alpha_0 \geq 2$, we derive

$$\begin{aligned}
|R_2(\alpha)| &\leq \frac{2}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2}} \sqrt{\frac{\overline{\alpha}}{\alpha_0}} \left(2 + \frac{49\sqrt{6}}{36} \frac{b_0}{b_0 - \overline{p}f}\right) \cdot \left| \int_0^{\kappa_2} e^{-\overline{\alpha}v} \mathrm{d}v + \int_0^{\kappa_1} e^{-\overline{\alpha}v} \mathrm{d}v \right| \\
&\leq \frac{1}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2}} \left(8 + \frac{49\sqrt{6}}{9} \frac{b_0}{b_0 - \overline{p}f}\right) \cdot \frac{1}{\sqrt{\overline{\alpha} \cdot \alpha_0}}.
\end{aligned}$$

Term R_1. The analysis of this term can be carried out as in Tiapkin et al. [2022] except the multiplication with $P(0)$,

$$|R_1(\alpha)| \leq \frac{c_1}{\sqrt{\sigma^2 c_\kappa \alpha_0} \cdot (1 - \lambda^\star(b_0 - u))} \cdot \frac{\exp(-c_\kappa \alpha_0)}{\overline{\alpha}^{1/2}},$$

where $c_1 = 2\sqrt{2}$.

Term R_3. We start from the bound

$$\begin{aligned}
\left| \int_K^\infty P(s) \exp(-\overline{\alpha}T(s)) \, \mathrm{d}s \right| &\leq \exp(-\overline{\alpha} \cdot \mathrm{Re}[T(K) - T(0)]) \cdot \exp(-\overline{\alpha}T(0)) \\
&\quad \cdot \sup_{s \in \mathbb{R}} |P(s)| \int_K^\infty \exp(-\overline{\alpha} \, \mathrm{Re}[T(s) - T(K)]) \, \mathrm{d}s.
\end{aligned}$$

Let us start from the analysis of an additional multiplier connected to $P(s)$

$$\sup_s |P(s)| = \sup_s \sqrt{\frac{1}{(1 - \lambda^\star(b_0 - u))^2 + s^2(b_0 - u)}} = \frac{1}{1 - \lambda^\star(b_0 - u)}.$$

The rest of the analysis coincides the the analysis of the same term in Tiapkin et al. [2022]

$$|R_3(\alpha)| \leq c_3 \cdot \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\mathrm{inf}}(\overline{p}, u, f))}{1 - \lambda^\star(b_0 - u)} \cdot \frac{1 - \lambda^\star(b_0 - u)}{b_0 - u} \exp(-c_\kappa \alpha_0)$$

for $c_3 = \sqrt{5\pi}/2$. □

Finally, we use a bounds on remainder terms to derive a lower bound on the density.

Lemma D.5. *Consider a function $f \in \mathrm{F}_m(b_0, b)$ and a vector $\alpha = (\alpha_0 + 1, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ with $\overline{\alpha} \geq 2\alpha_0$, $b_0 \geq 2b$. Let $w \sim \mathcal{D}\mathrm{ir}(\alpha)$ and assume that $Z = wf$ is non-degenerate. Let $\varepsilon \in (0, 1)$. Assume*

$$\alpha_0 \geq \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49 \cdot 4\sqrt{6}}{9}\right)^2 \frac{2}{\pi \cdot \varepsilon^2} + \log_{17/16}\left(\frac{5}{32 \cdot \varepsilon^2}\right) + \log_{17/16}(\overline{\alpha}).$$

Then for any $u \in (\overline{p}f, b_0)$,

$$p_Z(u) \geq (1 - \varepsilon)\sqrt{\frac{\overline{\alpha}}{2\pi}} \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\mathit{inf}}(\overline{p}, u, f))}{(1 - \lambda^\star(b_0 - u))\sqrt{\sigma^2}}.$$

Proof. We start the proof from the combination of Proposition D.3 and Proposition D.4

$$p_Z(u) \geq \frac{\overline{\alpha}}{2\pi}\left(\left(\sqrt{2\pi} - \frac{1}{\sqrt{\alpha_0}}\left(\frac{c_1}{\sqrt{c_\kappa}} + c_2\right)\right)\frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f))}{(1-\lambda^\star(b_0-u))\sqrt{\overline{\alpha}\cdot\sigma^2}} - |R_3(\alpha)|\right).$$

Since $\overline{\alpha} \geq 2\alpha_0$ and $b_0 \geq 2b$ we have $b_0/(b_0 - \overline{p}f) \leq 4$. In this case we have $c_\kappa \geq 1/2\log(17/16)$ and $c_2 \leq 8 + 49\sqrt{6}\cdot 4/9$. Therefore

$$\frac{c_1}{\sqrt{c_\kappa}} + c_2 \leq \frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49\cdot 4\sqrt{6}}{9} \triangleq \gamma_1.$$

And for $\alpha_0 \geq 4\gamma_1^2/(2\pi\cdot\varepsilon^2)$, we have

$$\begin{aligned} p_Z(u) &\geq \frac{\overline{\alpha}}{2\pi}\left(\sqrt{2\pi}(1-\varepsilon/2)\frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f)}{(1-\lambda^\star(b_0-u))\sqrt{\overline{\alpha}\cdot\sigma^2}} - |R_3(\alpha)|\right) \\ &\geq \frac{\sqrt{\overline{\alpha}}}{2\pi}\left(\frac{\sqrt{2\pi}(1-\varepsilon/2)}{\sqrt{\overline{\alpha}\,\sigma^2}} - c_3\cdot\frac{1-\lambda^\star(b_0-u)}{b_0-u}\cdot\exp(-c_\kappa\alpha_0)\right)\frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f)}{(1-\lambda^\star(b_0-u))}. \end{aligned}$$

Note that $\mathbb{E}[Y_0] = 0$ and observe that the inequality

$$\begin{aligned} \sigma^2 = \mathbb{E}[Y_0^2] = \mathrm{Var}[Y_0] &\leq \left(\frac{b_0-u}{2(1-\lambda^\star(b_0-u))} + \frac{u}{2(1+\lambda^\star u)}\right)^2 \\ &= \frac{b_0^2}{4(1-\lambda^\star(b_0-u))^2(1+\lambda^\star u)^2} \leq \frac{4(b_0-u)^2}{(1-\lambda^\star(b_0-u))^2}, \end{aligned}$$

follows from a general bound on variance of bounded random variables (bounded differences), the fact (see Lemma 12 in Honda and Takemura [2010])

$$\lambda^\star \geq \frac{u-\overline{p}f}{u(b_0-u)} \iff 1+\lambda^\star u \geq \frac{b_0-\overline{p}f}{b_0-u},$$

and the inequality $b_0/(b_0-\overline{p}f) \leq 4$. It yields

$$p_Z(u) \geq \frac{\sqrt{\overline{\alpha}}}{2\pi}\left(\sqrt{2\pi}(1-\varepsilon/2) - 2\sqrt{5\pi}\exp(-c_\kappa\alpha_0)\cdot\sqrt{\overline{\alpha}}\right)\frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f))}{(1-\lambda^\star(b_0-u))\cdot\sqrt{\sigma^2}}.$$

To guarantee

$$2\sqrt{5\pi}\exp(-c_\kappa\alpha_0)\cdot\sqrt{\overline{\alpha}} \leq \sqrt{2\pi}\cdot(\varepsilon/2)$$

we have to choose

$$\alpha_0 \geq \log_{17/16}(5/(32\varepsilon^2)) + \log_{17/16}(\overline{\alpha}).$$

It allows us to conclude

$$p_Z(u) \geq (1-\varepsilon)\sqrt{\frac{\overline{\alpha}}{2\pi}}\frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f))}{(1-\lambda^\star(b_0-u))\cdot\sqrt{\sigma^2}}.$$

□

Before proceeding with the final proof, we derive one important technical result.

Lemma D.6. *For any $u \in (\overline{p}f, b_0)$ it holds*

$$\mathcal{K}_{\mathit{inf}}(\overline{p},u,f) \geq \frac{1}{2}(\lambda^\star)^2\sigma^2\big(1-\lambda^\star(b_0-u)\big)^2.$$

Proof. Define the function $\phi_u(\lambda) = \mathbb{E}\log\big(1-\lambda(f(X)-u)\big)$ and $\lambda_u = \lambda^\star$. Remark that $\sigma^2 = -\phi_u''(\lambda_u)$. Thanks to the Taylor expansion of ϕ_u and the definition of λ_u it holds

$$0 = \phi_u(0) = \phi_u(\lambda_u) + 0 + \frac{\lambda_u^2}{2}\phi_u''(y\lambda_u)$$

for some $y \in (0,1)$. Thus we can rewrite $\mathcal{K}_{\text{inf}}$ as

$$\phi_u(\lambda) = \frac{\lambda_u^2}{2}(-\phi_u''(y\lambda_u))\,.$$

We will lower bound the opposite of the second derivative that appears above. First note that

$$-\phi_u''(y\lambda_u) = \mathbb{E}\left[\frac{(f(X)-u)^2}{\big(1-\lambda_u(f(X)-u)\big)^2}\left(\frac{1-\lambda_u(f(X)-u)}{1-y\lambda_u(f(X)-u)}\right)^2\right].$$

We now lower-bound the ratio, noting that if $X \le u$ then since $y \in (0,1)$

$$\frac{1-\lambda_u(f(X)-u)}{1-y\lambda_u(f(X)-u)} \ge 1\,.$$

In the other case $X > u$, we have $0 \le 1-y\lambda_u(f(X)-u) \le 1$ and $1-\lambda_u(f(X)-u) \ge 1-\lambda_u(b_0-u)$ thus

$$\frac{1-\lambda_u(f(X)-u)}{1-y\lambda_u(f(X)-u)} \ge 1-\lambda_u(b-u) > 0\,.$$

In both case using $1-\lambda_u(b_0-u) \le 1$ we get

$$\left(\frac{1-\lambda_u(f(X)-u)}{1-y\lambda_u(f(X)-u)}\right)^2 \ge \big(1-\lambda_u(b_0-u)\big)^2\,.$$

In particular, using the definition of $\phi''(u)$, it entails that

$$-\phi_u''(y\lambda_u) \ge -\phi_u''(\lambda_u)\big(1-\lambda_u(b-u)\big)^2\,.$$

Plugging this inequality in the integral representation of ϕ_u allows us to conclude

$$\phi_u(\lambda) \ge \frac{1}{2}\lambda_u^2\big(-\phi''(\lambda_u)\big)\big(1-\lambda_u(b-u)\big)^2\,.$$

□

Using this lemma we may proceed with the proof of our final result.

Proof of Theorem D.2. Define $Z = wf$. By Lemma D.5,

$$\mathbb{P}(wf \ge \mu) = \int_\mu^{b_0} p_Z(u)\mathrm{d}u \ge (1-\varepsilon)\sqrt{\frac{\overline{\alpha}}{2\pi}} \cdot \int_\mu^{b_0} \frac{\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},u,f))}{\sqrt{\sigma^2(1-\lambda^\star(b_0-u))^2}}\,\mathrm{d}u.$$

By Theorem 6 by Honda and Takemura [2010],

$$\frac{\partial}{\partial u}\mathcal{K}_{\text{inf}}(\overline{p},u,f) = \lambda^\star.$$

Thus, we can define a change of variables $t^2/2 = \mathcal{K}_{\text{inf}}(\overline{p},u,f)$, $t\mathrm{d}t = \lambda^\star\mathrm{d}u$ and write

$$\mathbb{P}(Z \ge \mu) = (1-\varepsilon)\int_{\sqrt{2\,\mathcal{K}_{\text{inf}}(\overline{p},\mu,f)}}^{+\infty} D(u)\sqrt{\frac{\overline{\alpha}}{2\pi}}\exp(-\overline{\alpha}t^2/2)\mathrm{d}t,$$

where $D(u)$ is defined as a positive square root of

$$D^2(u) = \frac{2\,\mathcal{K}_{\text{inf}}(\overline{p},u,f)}{(\lambda^\star)^2\sigma^2(1-\lambda^\star(b_0-u))^2}.$$

By Lemma D.6, $D^2(u) \ge 1$ and hence

$$\begin{aligned}\mathbb{P}(Z \ge \mu) &\ge (1-\varepsilon)\int_{\sqrt{2\,\mathcal{K}_{\text{inf}}(\overline{p},\mu,f)}}^{+\infty}\sqrt{\frac{\overline{\alpha}}{2\pi}}\exp(-\overline{\alpha}t^2/2)\mathrm{d}t \\ &= (1-\varepsilon)\mathbb{P}_{g\sim\mathcal{N}(0,1)}\big(g \ge \sqrt{2\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p},\mu,f)}\big).\end{aligned}$$

□

E Technical lemmas

E.1 On the Bernstein inequality

In this part, we restate Bernstein-type inequality of Talebi and Maillard [2018].

Lemma E.1 (Corollary 11 by Talebi and Maillard, 2018)**.** *Let $p, q \in \Delta_{S-1}$, where Δ_{S-1} denotes the probability simplex of dimension $S-1$. For all functions $f : \mathcal{S} \mapsto [0, b]$ defined on $\mathcal{S}$,*

$$
\begin{aligned}
pf - qf &\le \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p,q)} + \frac{2}{3} b\,\mathrm{KL}(p,q) \\
qf - pf &\le \sqrt{2\mathrm{Var}_q(f)\,\mathrm{KL}(p,q)}\,.
\end{aligned}
$$

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

Lemma E.2 (Lemma E.3 by Tiapkin et al., 2022)**.** *Let $p, q \in \Delta_{S-1}$ and a function $f : \mathcal{S} \mapsto [0, b]$, then*

$$
\begin{aligned}
\mathrm{Var}_q(f) &\le 2\mathrm{Var}_p(f) + 4b^2\,\mathrm{KL}(p,q)\,, \\
\mathrm{Var}_p(f) &\le 2\mathrm{Var}_q(f) + 4b^2\,\mathrm{KL}(p,q).
\end{aligned}
$$

Lemma E.3 (Lemma E.4 by Tiapkin et al., 2022)**.** *For $p, q \in \Delta_{S-1}$, for $f, g : \mathcal{S} \mapsto [0, b]$ two functions defined on $\mathcal{S}$, we have that*

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

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

F Lazy version of OPSRL

In this section we present Lazy-OPSRL a lazy version of the OPSRL algorithm. Following Efroni et al. [2019], instead of computing new Q-values by backward induction before each episode in Lazy-OPSRL we just just do one step of optimistic incremental planning at the current state to obtain improved Q-values (at the current state) and act greedily with respect to them. Precisely, given initial optimistic value functions $\overline{V}_h^{-1}(s) = r_0 H$ for all $(h,s) \in [H] \times \mathcal{S}'$ and Q-function $\overline{Q}_h^{-1}(s,a) = r_0 H$ for all $(h,s,a) \in [H] \times \mathcal{S}' \times \mathcal{A}$ we update Q-values by applying the Bellman operator *only at the visited states*:

$$\begin{aligned}
\overline{Q}_h^t(s,a) &\triangleq \mathbb{1}\{s = s_h^{t+1}\}\Big(r_h(s,a) + \max_{j\in[J]}\{\widetilde{p}_h^{t,j}\overline{V}_{h+1}^{t-1}(s,a)\}\Big) + (1 - \mathbb{1}\{s = s_h^{t+1}\})\overline{Q}_h^{t-1}(s,a)\,,\\
\overline{V}_h^t(s) &\triangleq \min\left\{\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a), \overline{V}_h^{t-1}(s)\right\},\\
\pi_h^{t+1}(s) &\in \arg\max_{a\in\mathcal{A}} \overline{Q}_h^t(s,a)\,,
\end{aligned} \tag{13}$$

where the posterior sample are still given by $\widetilde{p}_h^{t,j}(s,a) \sim \mathcal{D}\mathrm{ir}\Big(\big(\overline{n}_h^t(s'|s,a)/\kappa\big)_{s'\in\mathcal{S}'}\Big)$ and $\overline{V}_{H+1}^t(s) = 0$ for all t, s. Consequently Lazy-OPSRL enjoys a better time complexity of $\widetilde{\mathcal{O}}(HSA)$ per episode than the one $\widetilde{\mathcal{O}}(HS^2A)$ of OPSRL.

The complete description of Lazy-OPSRL is given in Algorithm 2 for a general family of probability distribution parameterized by the pseudo-counts over the transitions instead of the Dirichlet inflated prior/posterior.

Algorithm 2 Lazy-OPSRL

1: **Input:** Family of probability distributions $\rho : \mathbb{N}_+^{S+1} \to \Delta_{\mathcal{S}'}$ over transitions, initial pseudo-count $\overline{n}_h^0$, number of posterior samples J, initial value functions $\overline{V}_h^{-1}$, initial Q-functions $\overline{Q}_h^{-1}$.
2: **for** $t \in [T]$ **do**
3:   **for** $h \in [H]$ **do**
4:     Sample J independent transitions $\widetilde{p}_h^{t-1,j}(s,a) \sim \rho\big(\overline{n}_h^{t-1}(s'|s,a)_{s'\in\mathcal{S}'}\big)$, $\quad j \in [J]$.
5:     Compute for all $a \in \mathcal{A}$

$$\begin{aligned}
\overline{Q}_h^{t-1}(s_h^t,a) &= r_h(s_h^t,a) + \max_{j\in[J]}\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-2}(s_h^t,a)\}\,,\\
\overline{V}_h^{t-1}(s_h^t) &= \min\left\{\max_{a\in\mathcal{A}} \overline{Q}_h^{t-1}(s_h^t,a), \overline{V}_h^{t-2}(s_h^t)\right\}.
\end{aligned}$$

6:     Play $a_h^t \in \arg\max_{a\in\mathcal{A}} \overline{Q}_h^{t-1}(s_h^t,a)$.
7:     Observe $s_{h+1}^t \sim p_h(s_h^t,a_h^t)$.
8:     Increment the pseudo-count $\overline{n}_h^t(s_{h+1}^t|s_h^t,a_h^t)$.
9:   **end for**
10: **end for**

Interestingly, we can also obtain a regret bound for Lazy-OPSRL of the same order as OPSRL with the same number of posterior samples as in 3.1.

Theorem F.1. *Consider a parameter $\delta \in (0,1)$. Let $\kappa \triangleq 2(\log(12SAH/\delta) + 3\log(e\pi(2T+1)))$, $n_0 \triangleq \lceil\kappa(c_0 + \log_{17/16}(T))\rceil$, $r_0 \triangleq 2$, where c_0 is an absolute constant defined in (4); see Appendix B.2. Then for Lazy-OPSRL, with probability at least $1-\delta$,*

$$\mathfrak{R}^T = \mathcal{O}\Big(\sqrt{H^3SATL^3} + H^3S^2AL^3\Big),$$

where $L \triangleq \mathcal{O}(\log(HSAT/\delta))$.

Proof. Since this proof is very similar to the one of Theorem 3.1 we only describe how it needs to be adapted.

Optimism We are going to show that on event $\mathcal{E}^{\mathrm{anticonc}}(\delta)$ (see Proposition B.4 for definition) our estimate of Q-function is optimistic that is $\overline{Q}_h^t(s,a) \geq Q_h^\star(s,a)$ for any $(t,h,s,a) \in \{0,\ldots,T\} \times [H] \times \mathcal{S} \times \mathcal{A}$ and $\overline{V}_h^t(s) \geq V_h^\star(s)$ for $(t,h,s) \in \{-1,\ldots,T\} \times [H] \times \mathcal{S}$.

We prove by forward induction on t and backward induction on h. Base of induction $t=-1$ and $h = H+1$ is trivial. Next, if $s \neq s_h^{t+1}$ then $\overline{Q}_h^t(s,a) = \overline{Q}_h^{t-1}(s,a)$ and the statement is correct by the induction hypothesis. In the case of $s = s_h^{t+1}$ we have by Bellman equations and update rule (13)

$$
\overline{Q}_h^t(s,a) - Q_h^\star(s,a) = \max_{j\in[J]}\{\widetilde{p}_h^{t,j}\overline{V}_{h+1}^{t-1}(s,a)\} - p_h V_{h+1}^\star(s,a).
$$

By induction on t and h we have $\overline{V}_{h+1}^{t-1}(s') \geq V_{h+1}^\star(s')$ for any $s' \in \mathcal{S}$ thus by combination with event $\mathcal{E}^{\mathrm{anticonc}}(\delta)$ we conclude the statement.

Regret bound Recall $\delta_h^t = \overline{V}_h^{t-1}(s_h^t) - V_h^{\pi^t}(s_h^t)$ and $\overline{\mathfrak{R}}_h^T = \sum_{t=1}^t \delta_h^t$. By update rule for value function $\overline{V}_h^t(s_h^t) \leq \overline{Q}_h^t(s_h^t, a_h^t)$. Thus we can proceed by update rule for Q-function and Bellman equations

$$
\begin{aligned}
\delta_h^t &\leq \overline{Q}_h^{t-1}(s_h^t,a_h^t) - Q_h^{\pi^t}(s_h^t,a_h^t) = \max_{j\in[J]}\Big\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-2}(s_h^t,a_h^t)\Big\} - p_h V_{h+1}^{\pi^t}(s_h^t,a_h^t) \\
&= \underbrace{\max_{j\in[J]}\Big\{\widetilde{p}_h^{t-1,j}\overline{V}_{h+1}^{t-2}(s_h^t,a_h^t)\Big\} - \overline{p}_h^{t-1}\overline{V}_{h+1}^{t-2}(s_h^t,a_h^t)}_{(\mathbf{A})} + \underbrace{[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}_{h+1}^{t-2}(s_h^t,a_h^t)}_{(\mathbf{B})} \\
&+ \underbrace{[\widehat{p}_h^{t-1} - p_h][\overline{V}_{h+1}^{t-2} - V_{h+1}^\star](s_h^t,a_h^t)}_{(\mathbf{C})} + \underbrace{[\widehat{p}_h^{t-1} - p_h]V_{h+1}^\star(s_h^t,a_h^t)}_{(\mathbf{D})} \\
&+ \underbrace{p_h[\overline{V}_{h+1}^{t-2} - V_{h+1}^{\pi^t}](s_h^t,a_h^t) - [\overline{V}_{h+1}^{t-2} - V_{h+1}^{\pi^t}](s_{h+1}^t)}_{\xi_h^t} + \underbrace{[\overline{V}_{h+1}^{t-2} - \overline{V}_{h+1}^{t-1}](s_{h+1}^t)}_{\Delta_h^t} + \delta_h^t.
\end{aligned}
$$

Here we see that all terms are very similar to the terms that appears in the proof of Lemma B.8 except the additional one $\Delta_h^t \triangleq [\overline{V}_{h+1}^{t-2} - \overline{V}_{h+1}^{t-1}](s_{h+1}^t)$. By adapting the concentration event $\mathcal{G}^{\mathrm{conc}}(\delta)$ with a shift of indices we may obtain the following upper bound (for $N_h^t > 0$)

$$
\begin{aligned}
\delta_h^t &\leq \left(1 + \frac{1}{H}\right)\delta_h^t + \left(1 + \frac{1}{H}\right)\Delta_h^t + \left(1 + \frac{1}{H}\right)\xi_h^t \\
&+ 3L\sqrt{\frac{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-2}](s_h^t,a_h^t)}{\overline{N}_h^t}} + \sqrt{2L}\cdot\sqrt{\frac{\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t)}{N_h^t}} \\
&+ \frac{10H^2S\cdot L}{N_h^t} + \frac{16L^2H}{\overline{N}_h^t}.
\end{aligned}
$$

Thus, the surrogate regret is bounded by almost the same quantity up to a shift of indices and one additional term

$$
\overline{\mathfrak{R}}_h^T \leq \tilde{A}_h^T + B_h^T + C_h^T + 4\mathrm{e}H\sqrt{2HTL} + 2\mathrm{e}H^2SA + \sum_{t=1}^T\sum_{h'=h}^H \gamma_{h'}\Delta_{h'}^t,
$$

where $\gamma_h = (1+1/H)^{H-h+1}$ and

$$\tilde{A}_h^T = 3\mathrm{e}L \sum_{t=1}^{T}\sum_{h'=h}^{H} \sqrt{\mathrm{Var}_{\overline{p}_{h'}^{t-1}}[\overline{V}_{h+1}^{t-2}](s_{h'}^t, a_{h'}^t) \cdot \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}},$$

$$B_h^T = \mathrm{e}\sqrt{2L} \sum_{t=1}^{T}\sum_{h'=h}^{H} \sqrt{\mathrm{Var}_{p_{h'}}[V_{h+1}^\star](s_{h'}^t, a_{h'}^t) \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}},$$

$$C_h^T = 26H^2SL^2 \sum_{t=1}^{T}\sum_{h'=h}^{H} \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}.$$

The terms B_h^T and C_h^T remain exactly the same as in the analysis of OPSRL, whereas there will be a small difference in the analysis $\tilde{A}_h^T$.

Next, we analyze the new term using non-increasing of the value function $\overline{V}_h^{t-1}(s) \le \overline{V}_h^{t-2}(s)$

$$\sum_{t=1}^{T}\sum_{h'=h}^{H} \gamma_{h'} \Delta_{h'}^t \le \mathrm{e} \sum_{t=1}^{T}\sum_{h'=h}^{H} \Delta_{h'}^t .$$

We derive a bound on the sum of Δ_h^t over T for any fixed h by a telescoping property

$$\begin{aligned} \sum_{t=1}^{T} \Delta_h^t &= \sum_{s\in\mathcal{S}}\sum_{t=1}^{T} \mathbb{1}\{s = s_{h+1}^t\}[\overline{V}_{h+1}^{t-2} - \overline{V}_{h+1}^{t-1}](s) \\ &\le \sum_{s\in\mathcal{S}}\sum_{t=1}^{T} [\overline{V}_{h+1}^{t-2} - \overline{V}_{h+1}^{t-1}](s) = \sum_{s\in\mathcal{S}} [\overline{V}_{h+1}^{-1} - \overline{V}_{h+1}^{T-1}](s) \le 2HS. \end{aligned} \tag{14}$$

Thus we have a next bound for surrogate regret

$$\overline{\mathfrak{R}}_h^T \le \tilde{U}_h^T \triangleq \tilde{A}_h^T + B_h^T + C_h^T + 4\mathrm{e}H\sqrt{2HTL} + 4\mathrm{e}H^2SA.$$

Next we explain the analysis of term $\tilde{A}_1^T$. To do it, we analyze the sum of variance by following the step of Lemma B.10. All analysis remain exactly the same except the analysis of term $(\mathbf{Z})$, that can be handled by additional use of inequality (14)

$$\begin{aligned} (\mathbf{Z}) &= \sum_{t=1}^{T}\sum_{h=1}^{H} r_0 H p_h (\overline{V}_{h+1}^{t-2} - V_{h+1}^{\pi^t})(s_h^t, a_h^t) \\ &= 2H \sum_{t=1}^{T}\sum_{h=1}^{H} (\xi_h^t + \delta_h^t + \Delta_h^t) \le 4H^2\sqrt{2TL} + 2H^2\widetilde{U}_1^T + 2H^2S. \end{aligned}$$

The only difference is in the term $2H^2S$ that is a second-order term. Thus, the following version of Lemma B.10 holds for Lazy-OPSRL

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t, a_h^t)\mathbb{1}\{N_h^t > 0\} \le 2H^2T + 2H^2\tilde{U}_1^T + 40H^3S^2AL^3 + 32H^3S\sqrt{2ATL}$$

with the change only in a constant in front of the third term. The rest of the proof remains the same as in the analysis of OPSRL. □

G Experimental details

In this appendix we provides details on comparison OPSRL with some baselines and additionally study the impact of choice of the number of posterior samples J for PSRL and the impact of optimistic prior for OPSRL and PSRL. Our code is published on GitHub and based on the library rlberry by Domingues et al. [2021].

Environment We use a grid-world environment with 100 states $(i, j) \in [10] \times [10]$ and 4 actions (left, right, up and down). The horizon is set to $H = 50$. When taking an action, the agent moves in the corresponding direction with probability $1 - \varepsilon$, and moves to a neighbor state at random with probability ε. The agent starts at position $(1, 1)$. The reward equals to 1 at the state $(10, 10)$ and is zero elsewhere.

Number of posterior samples First we investigate the influence of the number of posterior samples J on the regret. We fixed the other parameters as follows: We use the prior over the transition probability described in Section 3 with $n_0 = 1$ initial pseudo-counts and no inflation $\kappa = 1$. In Figure 2 we plot the regret of OPSRL in the environment described above when the number of posterior samples varies in $J \in \{1, 4, 8, 16, 32\}$. We observe that the number of posterior samples has little effect on the regret, especially if we compare it to the scale of the gap between the different regret curves of the baselines in Figure 1. Thus, in the sequel of this appendix, we arbitrarily choose $J = 8$. Another justification of this choice is that $J \approx \log(T)$ for $T = 10000$, as it was required by theoretical analysis.

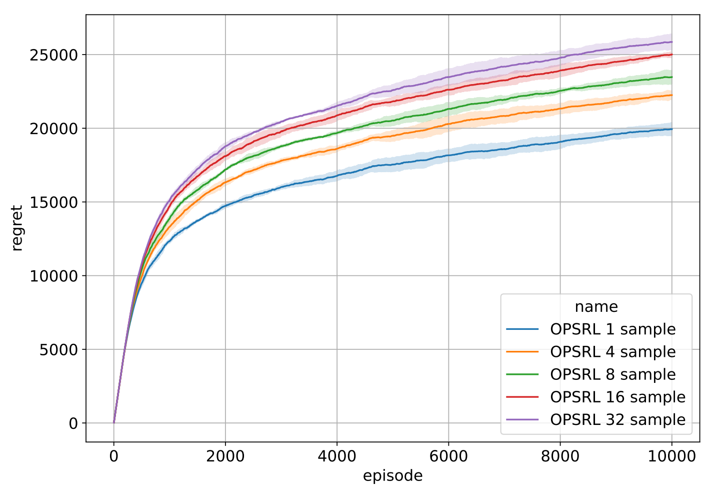

Figure 2: Regret of OPSRL for $J \in \{1, 4, 8, 16, 32\}$ for $H = 50$ and transitions noise 0.2. We show average over 4 seeds.

Baselines We compare OPSRL with the following baselines:

- The UCBVI algorithm by [Azar et al., 2017] (with Hoeffding-type bonuses). Since the theoretical bonus often leads to poor practical performance we use simplified bonuses from an idealized Hoeffding inequality of the form

$$\beta_h^t(s, a) = \min\left(\sqrt{\frac{(H - h + 1)^2}{4n_h^t(s, a)}}, H - h + 1\right).$$

- The UCBVI-B algorithm, the same algorithm as above but with simplified bonuses from an idealized Bernstein inequality:

$$\beta_h^t(s,a) = \min\left(\sqrt{\frac{\mathrm{Var}_{\widehat{p}^t}[V_{h+1}^{t-1}](s,a)}{n_h^t(s,a)}} + \frac{H-h+1}{n_h^t(s,a)}, H-h+1\right).$$

- The PSRL algorithm by [Osband et al., 2013]. For this algorithm we used a Dirichlet distribution of parameter $(1/S,\ldots,1/S)$ as prior on the transition probability.
- The RLSVI algorithm by [Osband et al., 2016b]. As for UCBVI we used a simplified variance for the Gaussian noise

$$\sigma_h^t(s,a) = \min\left(\sqrt{\frac{(H-h+1)^2}{4n_h^t(s,a)}}, H-h+1\right).$$

For the OPSRL we use the prior over the transition probability described in Section 1 with $n_0 = 1$ initial pseudo-counts and no inflation $\kappa = 1$. Note that the number of pseudo-counts is the same that for the one of the chosen prior for PSRL (where the sum of parameters is also one). As discussed above we pick $J = 8$ posterior samples.

Results In Figure 1, we plot the regret of the various baselines and OPSRL in the grid world environment. In this experiment, we observe that OPSRL achieves competitive results with respect to PSRL. It is not completely surprising since they share the same Bayesian model on the transitions up to the prior. We shall elaborate more on the influence of the prior below. We also note that OPSRL outperforms UCBVI and RLSVI. This difference may be explained by the fact that OPSRL's optimism implies (in the worst case) KL bonuses as in Filippi et al. [2010]. The KL bonuses are stronger than Bernstein bonuses, see Lemma E.1, because they somehow rely on all moments of the empirical distribution rather than the first two moments as in the case of Bernstein bonuses or first moments for Hoeffding bonuses or for the variance of the Gaussian noise in RLSVI. Note also that in OPSRL, we do not have to solve the complex convex program to compute the KL bonuses Filippi et al. [2010], which could be computationally intensive.

Influence of prior Next we study the influence of the prior for posterior sampling algorithms. Here we will refer to OPSRL as an optimistic prior choice and to PSRL as a uniform prior choice.

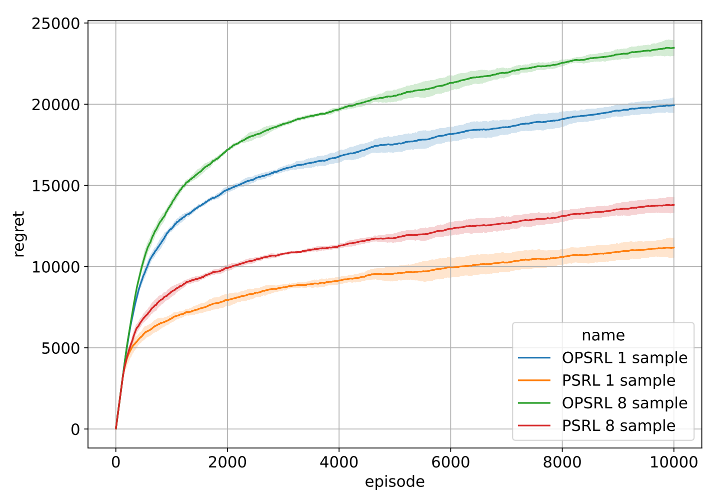

Figure 3: Regret of OPSRL with optimistic prior and PSRL with uniform prior for $J \in \{1, 8\}$ for $H = 50$ an transitions noise 0.2. We show average over 4 seeds.

On Figure 3 we may observe that algorithm convergences for both tested numbers of Thompson samples J and the only difference is the speed of forgetting the prior distribution that results in a constant difference between regrets. We see that optimistic prior is slightly harder to forget and it is connected to one of the most interesting features of it: optimistic prior is robust to the choice of the underlying probabilistic model. This property makes it universal at the price of efficiency on particular examples.