From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses

Daniil Tiapkin [1 2] **Denis Belomestny** [3 1] **Éric Moulines** [4 1] **Alexey Naumov** [1] **Sergey Samsonov** [1] **Yunhao Tang** [5] **Michal Valko** [5] **Pierre Ménard** [6]

Abstract

We propose the Bayes-UCBVI algorithm for reinforcement learning in tabular, stage-dependent, episodic Markov decision process: a natural extension of the Bayes-UCB algorithm by Kaufmann et al. (2012) for multi-armed bandits. Our method uses the quantile of a Q-value function posterior as upper confidence bound on the optimal Q-value function. For Bayes-UCBVI, we prove a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ where H is the length of one episode, S is the number of states, A the number of actions, T the number of episodes, that matches the lower-bound of $\Omega(\sqrt{H^3SAT})$ up to poly-log terms in H, S, A, T for a large enough T. To the best of our knowledge, this is the first algorithm that obtains an optimal dependence on the horizon H (and S) *without the need of an involved Bernstein-like bonus or noise.* Crucial to our analysis is a new fine-grained anti-concentration bound for a weighted Dirichlet sum that can be of independent interest. We then explain how Bayes-UCBVI can be easily extended beyond the tabular setting, exhibiting a strong link between our algorithm and Bayesian bootstrap (Rubin, 1981).

1. Introduction

In reinforcement learning (RL), an agent interacts with an environment with the objective of maximizing the sum of collected rewards (Puterman, 1994). In order to fulfill this objective, the agent should balance between *exploring the environment* and *exploiting the current knowledge* to accumulate rewards. In this paper aim at providing *generic solution* to this exploration-exploitation dilemma.

We model the environment as an unknown episodic tabular Markov decision process (MDP) with S states, A actions, and episodes of length H. After T episodes, we measure the performance of the agent by its cumulative regret which is the difference between the total reward collected by an optimal policy and the total reward collected by the agent during the learning. In particular, we study the *non-stationary* setting where rewards and transitions can change within an episode.

An effective and widely used way to solve the exploration-exploitation dilemma is the application of the principle of optimism in face of uncertainty. One line of work (Azar et al., 2017; Dann et al., 2017; Zanette and Brunskill, 2019a) for episodic MDPs and for non-episodic MDPs (Jaksch et al., 2010; Fruit et al., 2018; Talebi and Maillard, 2018) implements this principle by injecting optimism through *bonuses added to the rewards*. By adding these bonuses we can build upper confidence bounds (UCBs) on the optimal Q-value functions and act greedily with respect to them. Typically, these bonuses are decreasing functions of counts on the number of visits of state-action pairs. Notably, for such approach, Azar et al. (2017) proved a regret bound of order[1 2] $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$. Note that this upper bound matches, in the first order and up to poly-logarithmic terms, the known lower bound (Domingues et al., 2021b; Jin et al., 2018) of order $\Omega(\sqrt{H^3SAT})$ for the considered setting. The exploration based on building UCBs and adding bonuses is besides model-based also used for model-free algorithms (Jin et al., 2018; Zhang et al., 2020). For example, Zhang et al. (2020) proved a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ for an optimistic version of the Q-learning algorithm (Watkins and Dayan, 1992). One shortcoming of this exploration method is that algorithms with bonuses designed to obtain optimal problem-independent regret bound often perform poorly in practice, even for simple MDPs (Osband et al., 2013; Osband and Van Roy, 2017). Furthermore, the notion of count used in the bonuses does not easily generalize beyond the tabular setting[3] even if some solutions exist (Bellemare et al.,

[1]HSE University [2]Artificial Intelligence Research Institute [3]Duisburg-Essen University [4]École Polytechnique [5]DeepMind [6]Otto von Guericke University. Correspondence to: Daniil Tiapkin <dtyapkin@hse.ru>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

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

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

[3]Or simple linearly parameterized settings.

2016; Tang et al., 2017; Burda et al., 2019); See Section 4.1 for a thorough review of these methods.

A second line of work introduces optimism by *injecting noise*. Osband et al. (2013; 2016b); Osband and Van Roy (2017); Agrawal and Jia (2017) proposed the posterior sampling for RL (PSRL), an adaptation of the well-known Thompson sampling (Thompson, 1933) for multi-armed bandits. Using a Bayesian view, PSRL maintains a posterior on the MDP parameters and at each episode samples a new parameter from this posterior to act greedily with respect to it. Despite its good empirical performance in comparison to bonus-based algorithms (Osband et al., 2013; Osband and Van Roy, 2017), it is not known if PSRL algorithm can attain the problem-independent lower bound. Indeed, the best regret bound proved by Agrawal and Jia (2017); Qian et al. (2020) is of order $\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$ for PSRL. PSRL is close to the randomized least-square value iteration (RLSVI, Osband et al., 2016b; Russo, 2019) which injects noise directly in the value iteration through noisy Bellman updates. Specifically, a Gaussian with a variance that shrinks with the number of visits is added at each state-action pair during the value iteration. Interestingly, RLSVI also demonstrates good empirical performance in practice but most importantly can easily be extended outside the tabular setting as explained by Russo (2019); Osband et al. (2019), in particular, deep RL environment Osband et al. (2016a; 2018; 2019) Specifically, they combine RLSVI with DQN (Mnih et al., 2015) by replacing the Gaussian noise in RLSVI with a bootstrap sample (Efron, 1979) of the next targets. As a first step to analyze such noise, 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 adding pseudo rewards in the same fashion as Kveton et al. (2019). Their algorithm, BootNARL, comes with a regret bound of order $\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$. Note that Russo (2019) proved a regret bound of order $\widetilde{\mathcal{O}}(H^2S^{3/2}\sqrt{AT})$ for the original version of RLSVI. Later Xiong et al. (2021) improved this bound to $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ but at the price of scaling the Gaussian noise by a term similar to the Bernstein bonuses used in UCBVI. In particular, it is not clear if such variant could also be extended beyond the tabular setting.

Thus among the above Bayesian-inspired algorithms which both demonstrate good empirical results and are also readily extendable to large-scale environments none of them enjoys such as strong guarantee as problem-independent optimality. Therefore, in this paper, we propose to fill this gap with the Bayes-UCBVI algorithm. It is an optimistic algorithm that does not rely on bonuses but uses the *quantile of a of Q-value functions posterior* as UCBs on the optimal Q-value functions. We can think of Bayes-UCBVI as a deterministic version of PSRL, which, in particular, shares with PSRL the same good empirical performance, see Section 5. We adopt a *surrogate Bayesian model* for the transitions starting from an (improper) Dirichlet prior. *No assumption is made on the environment* and that the Bayesian model is purely instrumental for Bayes-UCBVI. The posterior on the Q-value function is then obtained by the Bellman equations. The prior can be interpreted as prior observations of pseudo-transitions toward an absorbing pseudo-state with maximal reward. As a result, Bayes-UCBVI has the advantage of requiring no information on the state space. We note that similar optimistic prior observations were already explored by Brafman and Tennenholtz (2002); Szita and Lőrincz (2008). For Bayes-UCBVI, we prove a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ matching the lower bound at first order and up to poly-log terms, see Table 1. In particular we get a tight dependence on the horizon H *without the need of an involved Bernstein-like bonus* (Azar et al., 2013; Zanette and Brunskill, 2019b; Zhang et al., 2020) *or Bernstein-type noise* (Xiong et al., 2021). Indeed, in Bayes-UCBVI the UCBs on the optimal Q-value functions induced by the Dirichlet posteriors over the transitions adapt to the variance automatically. Our proof relies on fine control of the deviations of the posterior. This tight control of the posterior is central in the analysis of the Bayesian inspired algorithm; see Agrawal and Jia (2017); Osband and Roy (2017). In particular, we provide a new anti-concentration inequality for a random Dirichlet-weighted sum that could be of independent interest, see Theorem 3.2. We believe that this anti-concentration inequality could also be used to tighten the bound of the PSRL algorithm.

As RLSVI, Bayes-UCBVI can be extended in a smooth way beyond the tabular setting. Indeed, we can reinterpret the posterior over the Q-value function of a given state-action pair as the distribution of a Bayesian bootstrap sample of the targets for this pair. Recall that in *Bayesian bootstrap* (Rubin, 1981) the observations are re-weighted by a Dirichlet sample *instead* of being sampled with replacement as done by the standard bootstrap (Efron, 1979). Consequently, the quantile serving as UCB can be straightforwardly approximated by Monte-Carlo method with Bayesian bootstrap samples. Thus, the exploration procedure of Bayes-UCBVI, can also be combined with the DQN algorithm to tackle large-scale RL: We achieve that by simply re-weighting the regression loss of the Q-value functions by a different Dirichlet sample. In particular, we explain how to combine the exploration procedure of Bayes-UCBVI with the DQN algorithm for Deep RL. The resulting algorithm is in essence an optimistic version of the one of Osband et al. (2019). We show experimentally that the resulting algorithm is competitive with the one introduced by Osband et al. (2019).

We highlight our main contributions:

- We propose the Bayes-UCBVI algorithm for tabular, stage-dependent, episodic RL. Interestingly Bayes-UCBVI is an optimistic algorithm that does not

Algorithm	Upper bound (non-stationary)
`UCBVI` (Azar et al., 2017) `UCB-Advantage` (Zhang et al., 2020) `RLSVI` (Xiong et al., 2021)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
`PSRL` (Agrawal and Jia, 2017) `BootNARL` (Pacchiano et al., 2021)	$\widetilde{\mathcal{O}}(H^2S\sqrt{AT})$
`Bayes-UCBVI` (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$
Lower bound (Jin et al., 2018; Domingues et al., 2021b)	$\Omega(\sqrt{H^3SAT})$

Table 1. Regret upper bound for episodic, non-stationary, tabular MDPs.

rely on adding bonuses but rather builds UCBs on the optimal Q-value functions by taking the quantile of a well-chosen posterior. For `Bayes-UCBVI`, we provide a regret bound of order $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ matching the problem independent lower bound up to poly-log terms.

- Central to the analysis of `Bayes-UCBVI` is a new anti-concentration inequality for a Dirichlet weighted sum (Theorem 3.2). We believe this inequality could be of independent interest, e.g., to sharpen the regret bound of other Bayesian inspired algorithms like `PSRL`.
- We extend `Bayes-UCBVI` beyond the tabular setting, exhibiting a strong link between our algorithm and Bayesian bootstrap (Rubin, 1981). In particular, we explain how to combine the exploration procedure of `Bayes-UCBVI` with the `DQN` algorithm for Deep RL. We show experimentally that the resulting algorithm is competitive with the one introduced by Osband et al. (2019).

2. Setting

We consider a finite episodic MDP $\big(\mathcal{S},\mathcal{A},H,\{p_h\}_{h\in[H]},\{r_h\}_{h\in[H]}\big)$, 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[4] 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 respectively optimal Bellman equations

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

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.[5] Before each episode t the agent select 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 $n_h^t(s,a) \triangleq \sum_{i=1}^t \mathbb{1}\big\{(s_h^i,a_h^i)=(s,a)\big\}$ are the number of times the state action-pair (s,a) was visited in step h in the first t episodes. Next, we define $n_h^t(s'|s,a) \triangleq \sum_{i=1}^t \mathbb{1}\big\{(s_h^i,a_h^i,s_{h+1}^i)=(s,a,s')\big\}$ the number of transitions from s to s' at step h.

Pseudo counts and improper Dirichlet distribution We define the pseudo counts as the counts shifted by initial pseudo counts $\overline{n}_h^t(s,a) \triangleq n_h^t(s,a)+n_0$. For $m\in\mathbb{N}^*$ the simplex of dimension $m-1$ is denoted by Δ_{m-1}. For $\alpha\in(\mathbb{R}_{++})^m$ we denote by $\mathcal{D}\mathrm{ir}(\alpha)$ the Dirichlet distribution on Δ_{m-1} with parameter α. We also extend this distribution to improper parameter $\alpha\in(\mathbb{R}_+)^m$ such that $\sum_{i=1}^m\alpha_i>0$ by injecting $\mathcal{D}\mathrm{ir}((\alpha_i)_{i:\alpha_i>0})$ into Δ_{m-1}. 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

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

[5] As explained by Fiechter (1994) and Kaufmann et al. (2021), 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 $p_{1'}(s_{1'},a)\triangleq p$.

over this set by $\mathcal{U}\mathrm{nif}[N]$. The vector of dimension N with all entries one is $\mathbf{1}^N \triangleq (1,\ldots,1)$. $\widehat{p}^{\,t}_h(s,a)$ is an empirical distribution defined as $\widehat{p}^{\,t}_h(s'|s,a) = n^t_h(s'|s,a)/n^t_h(s,a)$ if $n^t_h(s,a) > 0$ else $\widehat{p}^{\,t}_h(s'|s,a) = 1/S$, and $\overline{p}^{\,t}_h(s,a)$ is an pseudo-empirical measure defined as $\overline{p}^{\,t}_h(s,a) = \overline{n}^t_h(s'|s,a)/\overline{n}^t_h(s,a)$. Appendix A gives a reference of the notation used.

3. Algorithm

In this section we describe the `Bayes-UCBVI` algorithm. Similarly to `UCBVI`, we build upper confidence bounds (UCBs) on the Q-value functions and act greedily with respect to them. However, to construct the UCBs we instead use a quantile of certain posterior distribution. The name `Bayes-UCBVI` highlights the link between our algorithm and the one of Kaufmann et al. (2012) for multi-arm bandits.

First, 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 called "garden of even" was used by Brafman and Tennenholtz (2002); Szita and Lőrincz (2008). We denote the extended state space by $\mathcal{S}' \triangleq \mathcal{S} \cup \{s_0\}$. The optimal value at s_0 is $V^\star_h(s_0) = r_0(H-h+1)$ by definition. Next, we adopt a Bayesian model on the transition distributions. Note that it is only a surrogate model used by the algorithm but not the one from which the transition are sampled. Precisely, the improper prior on the probability transition is a Dirichlet[6] distribution $\rho^0_h(s,a) = \mathcal{D}\mathrm{ir}(\overline{n}^0_h(s'|s,a)_{s\in\mathcal{S}'})$ parameterized by the initial pseudo-count $\overline{n}^0_h(s'|s,a) = n_0\mathbb{1}\{s' = s_0\}$. We recall that the pseudo-counts $\overline{n}_h(s,a)$ are the counts plus a prior observation of a transition to the artificial state s_0. In fact, the prior is just a Dirac distribution at a deterministic transition $p_0(s') = \mathbb{1}_{\{s'=s_0\}}$ leading to the artificial state s_0. Then the posterior is a Dirichlet distribution $\rho^t_h(s,a) = \mathcal{D}\mathrm{ir}(\overline{n}^t_h(s'|s,a)_{s\in\mathcal{S}'})$. Given an upper bound on the value function at the next step $\overline{V}^t_{h+1}$, we set the upper confidence bound on the Q-value at step h to the quantile of order $\kappa^t_h(s,a)$ of the distribution of Q-value where the transition probability distribution is sampled according to the posterior,

$$\overline{Q}^t_h(s,a) \triangleq r_h(s,a) + \mathbb{Q}_{p\sim\rho^t_h(s,a)}\big(p\overline{V}^t_{h+1}(s,a), \kappa^t_h(s,a)\big),$$

where for $\rho \in \Delta_{S'-1}$, $\kappa \in [0,1]$, $V : \mathcal{S}' \to \mathbb{R}$, the quantile $\mathbb{Q}_{p\sim\rho}(pV,\kappa)$ of order κ is defined as $\mathbb{P}_{p\sim\rho}\big(pV \le \mathbb{Q}_{p\sim\rho}(pV,\kappa)\big) = \kappa$.

To compute UCBs on the value and Q-value functions for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we use an optimistic value iteration,

$$\begin{aligned}\overline{Q}^t_h(s,a) &= r_h(s,a) + \mathbb{Q}_{p\sim\rho^t_h(s,a)}\big(p\overline{V}^t_{h+1}(s,a), \kappa^t_h(s,a)\big)\\ \overline{V}^t_h(s) &\triangleq \max_{a\in\mathcal{A}} \overline{Q}^t_h(s,a) \qquad \overline{V}^t_h(s_0) \triangleq V^\star_h(s_0) \qquad (1)\\ \overline{V}^t_{H+1}(s) &\triangleq 0\,.\end{aligned}$$

The complete procedure of `Bayes-UCBVI` is described in Algorithm 1.

Algorithm 1 `Bayes-UCBVI`

1: **Input:** quantile functions $(\kappa^t)_{t\in[T]}$, prior dist. ρ^0
2: **for** $t \in [T]$ **do**
3: Optimistic planning, see (1)
4: **for** $h \in [H]$ **do**
5: Play $a^t_h \in \arg\max_{a\in\mathcal{A}} \overline{Q}^{t-1}_h(s^t_h, a)$
6: Observe $s^t_{h+1} \sim p_h(s^t_h, a^t_h)$
7: Update the posterior distributions $\rho^t_h(s^t_h, a^t_h)$ with $(s^t_h, a^t_h, s^t_{h+1})$
8: **end for**
9: **end for**

3.1. Analysis

We fix $\delta \in (0,1)$ and the quantile function

$$\kappa^t_h(s,a) \triangleq 1 - \frac{C_\kappa \delta}{SAH[2n^t_h(s,a)+1]^3[\overline{n}^t_h(s,a)]^{3/2}} \qquad (2)$$

up to an absolute constant $C_\kappa \triangleq 1/(5(\mathrm{e}\pi)^3)$. We now state the main result of the paper, proved in Appendix B.3 and with a proof sketch in Section 3.2.

Theorem 3.1. *Consider a parameter $\delta > 0$. Let $n_0 \triangleq \lceil c_{n_0} + \log_{17/16}(T)\rceil$, $r_0 \triangleq 2$, where c_{n_0} is an absolute constant defined in (4); see Appendix B.2. Then for* `Bayes-UCBVI`*, with probability at least $1-\delta$,*

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

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

Notice that `Bayes-UCBVI` matches the problem-independent lower bound $\Omega(\sqrt{H^3SAT})$ by Jin et al. (2018); Domingues et al. (2021b) for $T \ge H^3S^3A$ and up to poly-log terms in $H, S, A, T, 1/\delta$.

Complexity `Bayes-UCBVI` is a model-based algorithm, and thus gets the $\mathcal{O}(HS^2A)$ space complexity of `UCBVI`. Note that there is no closed-form solution to compute the quantile used in the UCB and thus we approximate it e.g., by Monte-Carlo sampling; see Section 4. In particular, if we use B Monte-Carlo samples to approximate one quantile the time complexity of `Bayes-UCBVI` is $\mathcal{O}(BHS^2AT)$ for T episodes.

[6] See Section 2 for the extension of a Dirichlet distribution to parameter with coordinates that could be equal to zero.

Comparison with PSRL and RLSVI Bayes-UCBVI is close to PSRL (Osband et al., 2013; Agrawal and Jia, 2017). Instead of computing quantiles, PSRL directly samples a transition probability distribution from the posterior to compute Q-values. Note that these Q-values may not necessary be UCBs as for Bayes-UCBVI. Agrawal and Jia (2017) proved a regret bound of order $\widetilde{\mathcal{O}}(H^2 S\sqrt{AT})$ for PSRL. We believe that our analysis for Bayes-UCBVI and in particular Theorem 3.2, could be used to improve the regret bound for PSRL to $\widetilde{\mathcal{O}}(\sqrt{H^3 SAT})$, thus matching (in terms of its dependence on the number of states S and the horizon H) the *Bayesian* regret bound provided by Osband and Van Roy (2017). Another Bayesian-inspired algorithm is RLSVI by Osband et al. (2013). It works by injecting Gaussian noise into the Bellman equations. Adding this noise can be seen as sampling accordingly to a certain posterior on the Q-value functions (Russo and Van Roy, 2014). Recently, Xiong et al. (2021) improved the dependence on the horizon H in RLSVI's regret bound to $\widetilde{\mathcal{O}}(\sqrt{H^3 SAT})$ thanks to a Gaussian noise with a "Bernstein" shaped variance. Yet, it is not clear if this variant of RLSVI has a clean extension beyond the tabular setting. The interesting property of Bayes-UCBVI is that its Dirichlet posterior on the transitions *adjusts automatically to the variance without the need of Bernstein bonuses/noises*. Moreover, Pacchiano et al. (2021) proposed to replace the Gaussian noise in RLSVI by bootstrap sampling of past rewards and adding pseudo-rewards as Kveton et al. (2019). They proved[7] a regret bound of order $\widetilde{\mathcal{O}}(H^2 S\sqrt{AT})$ for this type of noise. Note that in Bayes-UCBVI it is targets rather than the rewards that are used in the (Bayesian) bootstrap, see Section 4.

3.2. Proof sketch

We now sketch the proof of Theorem 3.1. The proof relies heavily on *boundary-crossing probabilities for weighted sums of the Dirichlet distribution with integer parameter*. The result below gives tight bounds for these probabilities. The lower bound in particular, is one of our main technical contributions.

Step 1. Dirichlet boundary crossing

Theorem 3.2 (see Lemma D.1 and Theorem D.2). *For any $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{N}^{m+1}$ define $\overline{p} \in \Delta_m$ with $\overline{p}(\ell) = \alpha_l/\overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Assume that $\alpha_0 \geq \log_{17/16}(\overline{\alpha}) + c_{n_0}$, where c_{n_0} is defined in (4); see Appendix B.2, and $\overline{\alpha} \geq 2\alpha_0$. Then for any $f\colon \{0, \ldots, m\} \to [0, b_0]$ such that $f(0) = b_0$, $f(\ell) \leq b < b_0/2$, $\ell \in \{1, \ldots, m\}$ and any $\mu \in (\overline{p}f, b_0)$ we have*

$$\frac{\mathrm{e}^{-\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p},\mu,f)}}{\overline{\alpha}^{3/2}} \leq \mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \leq \mathrm{e}^{-\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p},\mu,f)},$$

where $\mathcal{K}_{\mathit{inf}}(p, u, f)$ is given by

$$\mathcal{K}_{\mathit{inf}}(p, u, f) \triangleq \max_{\lambda\in[0,1]} \mathbb{E}_{X\sim p}\left[\log\left(1 - \lambda\frac{f(X) - u}{b_0 - u}\right)\right].$$

While the upper bound follows directly from the work of Riou and Honda (2020), the lower bound is new. The proof of the lower bound is presented in Theorem D.2 and consists of two main steps:

1. Geometrical reduction of the density of wf to 1D complex integral (see Dirksen, 2015; Lasserre, 2020);
2. Sharp non-asymptotic analysis of the integral using the saddle-point method (see Olver, 1997; Fedoryuk, 1977).

Using Theorem 3.2, we show that $\overline{Q}^t$ is an upper confidence bound on the optimal action-value function.

Step 2. Optimism Using the lower bound of Theorem 3.2, we show that for our choice of $\kappa_h^t(s,a)$, given in (2), the following result holds.

Lemma 3.3 (see Lemma B.5). *Let $n_0 \geq \log_{17/16}(\overline{\alpha}) + c_{n_0}$ and $r_0 \geq 2$, where c_{n_0} is defined in (4); see Appendix B.2. Then on event $\mathcal{E}^\star(\delta)$; see Appendix C, for any $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$\mathbb{Q}_{p\sim\rho_h^t(s,a)}\big(pV^\star_{h+1}(s,a), \kappa_h^t(s,a)\big) \geq p_h V^\star_{h+1}(s,a).$$

By the decomposition (1) and the Bellman equation, we see that

$$\begin{aligned}\overline{Q}_h^t(s,a) &- Q_h^\star(s,a)\\ &\geq \mathbb{Q}_{p\sim\rho_h^t(s,a)}\big(p\overline{V}^t_{h+1}(s,a), \kappa_h^t(s,a)\big) - p_h V^\star_{h+1}(s,a).\end{aligned}$$

Induction over h and Lemma 3.3 yield that on event $\mathcal{E}^\star(\delta)$, $\overline{Q}_h^t(s,a) \geq Q_h^\star(s,a)$ for any $t \leq T, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$.

Step 3. Reduction to UCBVI with Bernstein bonuses By optimism we have

$$\mathfrak{R}^T \triangleq \sum_{t=1}^T V_1^\star(s_1^t) - V_1^{\pi_t}(s_1^t) \leq \sum_{t=1}^T \delta_1^t,$$

where $\delta_h^t \triangleq \overline{V}_h^{t-1}(s_1^t) - V_h^{\pi_t}(s_1^t)$. The quantity δ_h^t can be decomposed as follows

$$\begin{aligned}\delta_h^t = &\underbrace{\mathbb{Q}_{p\sim\rho_h^{t-1}(s_h^t,a_h^t)}(p\overline{V}^{t-1}_{h+1}(s_h^t,a_h^t), \hat{\kappa}_h^t) - \overline{p}_h^{t-1}\overline{V}^{t-1}_{h+1}(s_h^t,a_h^t)}_{(\mathbf{A})}\\ &+ \underbrace{[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}^{t-1}_{h+1}(s_h^t,a_h^t)}_{(\mathbf{B})}\\ &+ \underbrace{(\widehat{p}_h^{t-1} - p_h)[\overline{V}^{t-1}_{h+1} - V^\star_{h+1}](s_h^t,a_h^t)}_{(\mathbf{C})} + \underbrace{(\widehat{p}_h^{t-1} - p_h)V^\star_{h+1}(s_h^t,a_h^t)}_{(\mathbf{D})}\\ &+ \underbrace{p_h[\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_h^t,a_h^t) - [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s^t_{h+1})}_{\xi_h^t} + \delta^t_{h+1},\end{aligned}$$

[7] We hypothesise that the noise should be scaled by H as in RLSVI for their result to be valid.

where $\hat{\kappa}_h^t = \kappa_h^{t-1}(s_h^t, a_h^t)$. The terms $(\mathbf{C}), (\mathbf{D})$ and ξ_h^t coincide with similar terms in the analysis of UCBVI with Bernstein bonuses. The term $(\mathbf{B})$ could be upper-bounded by $\frac{r_0 H}{\max\{n_h^{t-1}, 1\}}$ and turns out to be a second-order term. The analysis of the term $(\mathbf{A})$ is novel. Using the upper bound from Theorem 3.2 we may obtain the Bernstein type inequality for the Dirichlet distribution (see Lemma C.8 in Appendix C) which yields the following key inequality for the quantile $\mathbb{Q}_{p\sim\rho_h^t(s,a)}(p\overline{V}_{h+1}^t(s,a), \kappa_h^t(s,a))$.

Lemma 3.4 (see Corollary B.7)**.** *Assume conditions of Theorem 3.1. On event $\mathcal{E}^\star(\delta)$, for any $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$\begin{aligned}
&\mathbb{Q}_{p\sim\rho_h^t(s,a)}(p\overline{V}_{h+1}^t(s,a), \kappa_h^t(s,a)) \\
&\leq \overline{p}_h^t \overline{V}_{h+1}^t(s,a) + 2\sqrt{\operatorname{Var}_{\overline{p}_h^t}[\overline{V}_{h+1}^t](s,a)\frac{\log\left(\frac{1}{1-\kappa_h^t(s,a)}\right)}{\overline{n}_h^t(s,a)}} \\
&\quad + \frac{2\sqrt{2}\cdot r_0 H \log\left(\frac{1}{1-\kappa_h^t(s,a)}\right)}{\overline{n}_h^t(s,a)}.
\end{aligned}$$

Since $1-\kappa_h^t(s,a)$ depends on $n_h^t(s,a)$ only polynomially, we see that the term $(\mathbf{A})$ can be upper bounded by a quantity which looks very similar to the Bernstein bonuses in UCBVI and, moreover, it has the same role in the regret analysis. After using these upper bounds, the rest of the proof follows from the analysis of UCBVI with the Bernstein bonuses; see Azar et al., 2017.

4. Bayes-UCBVI for Deep RL

We now extend Bayes-UCBVI beyond the tabular setting. Fix a state-action pair (s,a). At episode t, the targets to estimate the Q-value function at state-action pair (s,a) at step h are $y_h^n(s,a) \triangleq r_h(s,a) + \overline{V}_{h+1}^t(s_{h+1}^n)$ for $n \in [n_h^t(s,a)]$ where s_{h+1}^n is the next state observed after taking the action a in state s for the n^{th} time.[8] We also need prior targets[9] $y_h^n(s,a) \triangleq r_h(s,a) + \overline{V}_h^t(s_0)$ for $(-n+1) \in [n_0]$ corresponding to the pseudo-transition to s_0. Using the *aggregation property of the Dirichlet distribution* we can compute the UCB by taking the quantile of randomly re-weighted sum of targets. Precisely, we have that

$$\begin{aligned}
\overline{Q}_h^t(s,a) &\triangleq r_h(s,a) + \mathbb{Q}_{p\sim\rho_h^t(s,a)}(p\overline{V}_{h+1}^t(s,a), \kappa_h^t(s,a)) \\
&= \mathbb{Q}_{w\sim\mathcal{D}\mathrm{ir}(\mathbf{1}^{\overline{n}_h^t(s,a)})}\left(\sum_{n=-n_0+1}^{n_h^t(s,a)} w_n y_h^n(s,a), \kappa_h^t(s,a)\right).
\end{aligned}$$

We can approximate this quantile by the quantile of the empirical distribution of B Bayesian bootstrap samples (Rubin, 1981). Precisely, if we fix $(w_h^b(s,a))_{b\in[B]}$ i.i.d. samples from a Dirichlet distribution $\mathcal{D}\mathrm{ir}(\mathbf{1}^{\overline{n}_h^t(s,a)})$ we have

$$\begin{aligned}
&\overline{Q}_h^t(s,a) \approx \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}([B])}\left(\overline{Q}_h^{t,b}(s,a), \kappa_h^t(s,a)\right) \\
&\qquad \text{where } \overline{Q}_h^{t,b}(s,a) \triangleq \sum_{n=-n_0+1}^{n_h^t(s,a)} w_h^{n,b}(s,a) y_h^n(s,a)\,.
\end{aligned}$$

In particular, using that a uniform Dirichlet distribution can be obtained by normalizing independent samples from the exponential probability distribution $\mathcal{E}(1)$, we can obtain the Bayesian samples by solving a weighted linear regression

$$\overline{Q}_h^{t,b}(s,a) = \arg\min_x \sum_{n=-n_0+1}^{n_h^t(s,a)} z_h^{n,b}(s,a)(x - y_h^n(s,a))^2 \tag{3}$$

$$\text{where } z_h^{n,b}(s,a) \sim \mathcal{E}(1) \text{ i.i.d.}\,.$$

We name this approximation of Bayes-UCBVI, *incremental Bayes-UCBVI* (Incr-Bayes-UCBVI) and provide its detailed pseudo-code as Algorithm 2 in Appendix F. Note that this way to generate bootstrap sample is similar to the incremental Bayesian bootstrap by Osband and Van Roy, 2015 (their Algorithm 5).

A great advantage of this formulation of Bayes-UCBVI is that it can be easily extended beyond the tabular setting. Indeed, we can simply replace the weighted linear regression loss in (3) by the weighted regression loss of any function approximation. Remarkably, except for the initial pseudo-transitions, Incr-Bayes-UCBVI does not rely on counts but on a easy-to-implement Bayesian bootstrap. As an example, in Appendix F, we combine the Incr-Bayes-UCBVI exploration procedure with DQN (Mnih et al., 2015) and call it Bayes-UCBDQN, detailed as Algorithm 3 of Appendix F.

4.1. Related work

Generalizing principled solutions of the exploration-exploitation dilemma from the theoretical tabular RL setting to large-scale deep RL is quite challenging (Yang et al., 2021). For instance, Bellemare et al. (2016); Ostrovski et al. (2017) approach the count-based UCBs used in tabular RL by approximating the visits counts using a density estimation. Later, Tang et al. (2017) directly map states to hash codes and then count their occurrences in a hash table. Another line of work sets bonuses to the approximation error of some quantities related to the MDP dynamics: the forward dynamics (Schmidhuber, 1991; Pathak et al., 2017), the inverse dynamics (Haber et al., 2018) or simply a constant function (Fox et al., 2018). Similarly, Burda et al. (2019) builds bonus from he prediction error of a randomly initialized network. This can be further combined with the pseudo-counts (Badia et al., 2020) leading to impressive

[8]In particular, n is a number of visits of a state-action pair (s,a) and not the global time (the number of episodes).

[9]In the case of unknown rewards we use the sample returns instead. For the pseudo-target, we always set the rewards to 1 which gives $y_h^0(s,a) \triangleq H - h + 1$.

results. As in the tabular setting, a second line of work deals with the exploration-exploitation trade-off by injecting noise. Fortunato et al. (2018) add parametric noise to the weights of the agent's network that is learned with the weights. Azizzadenesheli et al. (2018) approximate PSRL by replacing the typical last linear layer of agent's Q-value network with a Bayesian linear regression. Alternatively, bootstrap DQN (BootDQN, Osband et al., 2016a; 2018; 2019) extends RLSVI by bootstrap sampling of the transitions to inject noise into DQN. Specifically, in BootDQN an ensemble of Q-value functions is learned each on a *different* bootstrap sample of the transitions collected so far. Building on this work, Nikolov et al. (2019) use bootstrap as well but instead combines the bandit algorithm, information direct sampling (Russo and Van Roy, 2014), with DQN. Recently, Bai et al. (2021) also proposed an optimistic algorithm based on bootstrap, using a bonus that scales with the variance of the ensemble of Q-value functions learned as did Osband et al. (2019).

Further comparison of Bayes-UCBDQN with BootDQN Bayes-UCBDQN is close to BootDQN of Osband et al. (2016a). In BootDQN, an ensemble of B bootstrap Q-value functions (or in practice only the "heads" of a unique Q-value function) are learned with different sub-sets of transitions. Each transition is randomly assigned to the training of one bootstrap Q-value function with a fixed probability $p \in [0, 1]$. In particular they consider $p = 0.5$ for the double-or-nothing bootstrap and $p = 1$ for no bootstrap. Each bootstrap Q-value function is then trained with targets computed from the *corresponding* bootstrap Q-value function at the next state; see Appendix G for a detailed description. As explained by Osband et al. (2016a), BootDQN can be seen as approximation of RLSVI. That is why Bayes-UCBDQN can be seen as an optimistic version of BootDQN (as Bayes-UCBVI is an optimistic version of PSRL & RLSVI). The main differences between BootDQN and Bayes-UCBDQN are therefore: (i) Bayes-UCBDQN acts greedily with respect to the quantile of the bootstrap Q-value functions instead of one bootstrap Q-value function sampled uniformly at random. (ii) Bayes-UCBDQN uses Bayesian bootstrap instead of the classical bootstrap (Efron, 1979). (iii) In Bayes-UCBDQN, the bootstrap Q-value functions are trained with the same target computed with the quantile of the bootstrap Q-functions at the next step, as in (3). We discuss the impacts of these modifications in Section 5.

5. Experimental Results

In this section we provide experiments on Bayes-UCBVI and its variants. We illustrate two points: First, that Incr-Bayes-UCBVI performs similarly as other algorithms relying on noise-injection for exploration such that PSRL and RLSVI. Second, that Bayes-UCBDQN, the deep RL extension of Bayes-UCBVI is competitive with BootDQN.

5.1. Tabular environment

We first evaluate Bayes-UCBVI and Incr-Bayes-UCBVI on a simple tabular environment.

Environment For the tabular experiments we consider a simple grid-world with 5 connected rooms of size 5×5, totalling $S = 129$ states. The agent starts in the middle room. There is one small deterministic reward in the leftmost room, one large deterministic reward in the rightmost room and zero reward elsewhere. The agent can take $A = 4$ actions: moving up, down, left, right. When taking an action, the agent moves in the corresponding direction with probability 0.9 and moves to a neighboring state at random with probability 0.1. The horizon is fixed to $H = 30$; see Appendix G for details. In this environment the agent must explore efficiently all the room avoiding being lured by the small reward in the leftmost room.

Baselines We compare Bayes-UCBVI and Incr-Bayes-UCBVI with the following baselines: UCBVI (Azar et al., 2017), RLSVI (Osband et al., 2016b), and PSRL (Osband et al., 2013); see Appendix G for a full description of the parameters of the algorithms used in the experiments.

Results In Figure 1, we plot the regret of the various baselines, Bayes-UCBVI and Incr-Bayes-UCBVI in the aforementioned environment. In this experiment, we observe that both Bayes-UCBVI and Incr-Bayes-UCBVI achieve competitive results with respect to baselines relying on noise-injection for exploration (PSRL, RLSVI). This is remarkable, since the *common belief is that optimistic algorithm perform poorly in practice* (Osband and Van Roy, 2017). Indeed, Incr-Bayes-UCBVI exhibits a regret similar to PSRL. It is not completely surprising since they share the same model on the transitions (up to the prior). Notice that Bayes-UCBVI performs slightly worse than Incr-Bayes-UCBVI but better than RLSVI. One possible reason to explain this gap between Bayes-UCBVI and Incr-Bayes-UCBVI is that the incremental implementation of Bayesian bootstrap forgets the prior faster than the non-incremental version, resulting in a more aggressive algorithm. We also note that RLSVI performs slightly worse than PSRL, Incr-Bayes-UCBVI but much better than UCBVI. A possible explanation for this ranking is that RLSVI is much more aggressive than UCBVI when they have comparable noise, bonuses; whereas PSRL, Incr-Bayes-UCBVI, Bayes-UCBVI take better advantage of the small variance of this particular environment than the two last baselines.

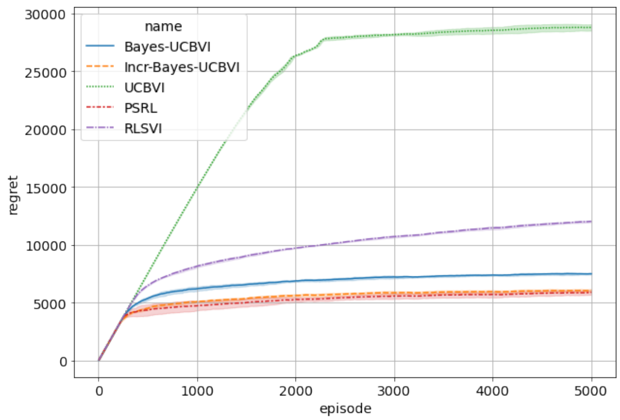

Figure 1. Regret of `Bayes-UCBVI` and `Incr-Bayes-UCBVI` compared to baselines for $H = 30$ an transitions noise 0.1. We show average over 4 seeds.

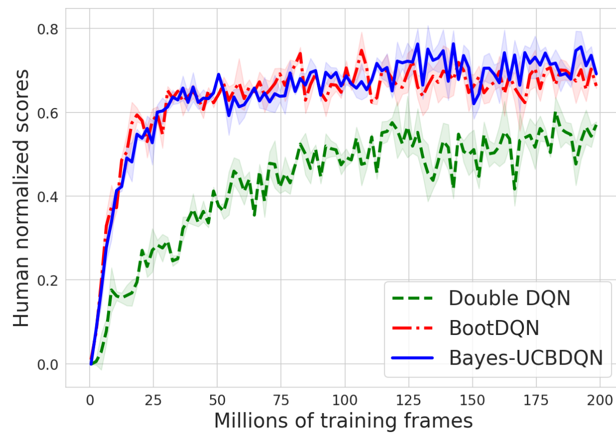

Figure 2. Evaluating deep RL algorithms with median human normalized scores across Atari-57 games. We compare `DoubleDQN`, `BootDQN` and `Bayes-UCBDQN`. The training curves show the average $\pm$ std over 3 seeds.

5.2. Deep RL experiments

In this section we evaluate the performance of `Bayes-UCBDQN` in large-scale environments.

Setup All algorithms are based on the architecture of `DQN` (Mnih et al., 2013). In order to implement the bootstrapped ensemble, we follow `BootDQN` (Osband and Van Roy, 2015) and maintain an ensemble of $B = 10$ head networks over a shared torso network. For fairness of comparison, all algorithmic variants share hyper-parameters wherever possible; see Appendix G for further details on the detailed architecture and implementation details.

Environment and evaluation To evaluate the scalability of `Bayes-UCBDQN`, we train DQN variants over a suite of 57 Atari games (Bellemare et al., 2013). For each algorithm and each game, we train for 200M frames and record the human normalized scores per game. The overall performance curve in Figure 2 is calculated as the median score over all games.

Results We compare `DoubleDQN` (Van Hasselt et al., 2016), `BootDQN` and `Bayes-UCBDQN`. In Figure 2, we show the evaluation performance of different algorithms over training, measured in median human normalized scores. We make a few observations: (1) Both `Bayes-UCBDQN` and `BootDQN` outperform `DoubleDQN`, potentially due to better training stability thanks to more consistent exploration; (2) The performance of `BootDQN` converges to about 0.7, which is consistent with results of Osband and Van Roy (2015); (3) Overall, `Bayes-UCBDQN` and `BootDQN` perform similarly. We see that `Bayes-UCBDQN` achieves very marginal advantage over `BootDQN` towards the end of training, however, more significant gains might require further engineering efforts. Nevertheless, we have established that `Bayes-UCBDQN`, as a theoretically grounded algorithm, is competitive with `BootDQN`. This paves the way for future research in this space; see Appendix G for further discussion on the effect of various hyper-parameters.

6. Conclusion

We presented a new algorithm, `Bayes-UCBVI`. It is an optimistic algorithm that does not rely on bonuses but rather uses the quantile of a well-chosen posterior to inject optimism. We proved that this algorithm is problem-independent optimal up to term poly-log in the horizon H, the number of action A, states S and episodes T. `Bayes-UCBVI` also exhibits similar empirical performance than other existing Bayesian-inspired algorithms thus bridging the optimal problem-independent theoretical guarantees of optimistic algorithms and the good empirical results of algorithms relying on noise-injection for exploration. Importantly we also demonstrated that `Bayes-UCBVI` could easily be extended beyond the tabular setting. In particular, we provided a new principled algorithm `Bayes-UCBDQN` based on `Bayes-UCBVI` that is competitive with `BootDQN` of Osband et al. (2019) on large-scale environments.

This work also raises the following open question that we think bring interesting future directions.

Problem-independent optimality of `PSRL` Central to the proof of the regret bound of `PSRL` is the control of the deviation of a Dirichlet re-weighted sum (Agrawal and Jia, 2017). Thus, we believe that the anti-concentration inequality of Theorem 3.2, or a close variant, could allow to improve the regret bound to $\widetilde{\mathcal{O}}(\sqrt{H^3SAT})$ (for T large enough). In particular, this would imply that `PSRL` is also problem-independent optimal.

Integration with deep RL agents Existing deep RL architectures, such as the implementations of base agent's loss functions and training pipeline, might not interact with our proposed exploration techniques in the optimal way (see Appendix G.3 for details). Thus, an open question is whether we could make more fundamental changes to certain deep RL agents so that the exploration methods can be integrated in a better way.

Acknowledgements

D. Belomestny acknowledges the financial support from Deutsche Forschungsgemeinschaft (DFG), Grant Nr. 497300407. P. Ménard is supported by the SFI Sachsen-Anhalt for the project RE-BCI ZS/2019/10/102024 by the Investitionsbank Sachsen-Anhalt. The work of D. Tiapkin, A. Naumov, and S. Samsonov were supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

References

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., 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3):325–349, 2013.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop, ITA 2018, San Diego, CA, USA, February 11-16, 2018*, pages 1–9. IEEE, 2018.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, and Charles Blundell. Never Give Up: Learning directed exploration Strategies. In *International Conference on Learning Representations*, 2020.

Chenjia Bai, Lingxiao Wang, Lei Han, Jianye Hao, Animesh Garg, Peng Liu, and Zhaoran Wang. Principled exploration via optimistic bootstrapping and backward induction. 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 577–587. PMLR, 2021.

Joan Bas-Serrano, Sebastian Curi, Andreas Krause, and Gergely Neu. Logistic q-learning. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3610–3618. PMLR, 2021.

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

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Jonathan (Jon) Borwein and O-Yeat Chan. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12, 2007.

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.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In *Neural Information Processing Systems*, 2017.

Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *Annals of probability*, 32:1902–1933, 2004.

Hauke Carl-Erwin Dirksen. *Sections of simplices and cylinders: Volume formulas and estimates*. PhD thesis, 2015.

Omar Darwiche Domingues, Yannis Flet-Berliac, Edouard Leurent, Pierre Ménard, Xuedong Shang, and Michal Valko. rlberry - A Reinforcement Learning Library for Research and Education, 2021a.

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited. In *Algorithmic Learning Theory*, 2021b.

Omar Darwiche Domingues, Pierre Menard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Kernel-based reinforcement learning: A finite-time analysis. 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 2783–2792. PMLR, 2021c.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979.

Lawrence C Evans and Ronald F Garzepy. *Measure theory and fine properties of functions*. Routledge, 2018.

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

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Conference on Learning Theory*, 1994.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Rémi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Lior Fox, Leshem Choshen, and Yonatan Loewenstein. DORA the explorer: Directed outreaching reinforcement action-selection. In *International Conference on Learning Representations*, 2018.

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 and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In Sham M. Kakade and Ulrike von Luxburg, editors, *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 359–376, Budapest, Hungary, 2011. PMLR.

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.

Richard A. Groeneveld and Glen Meeden. The mode, median, and mean inequality. *The American Statistician*, 31 (3):120–121, 1977.

Nick Haber, Damian Mrowca, Stephanie Wang, Li F Fei-Fei, and Daniel L Yamins. Learning to play with intrinsically-motivated, self-aware agents. 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.

Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23:2613–2621, 2010.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

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.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 99:1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sébastien Bubeck, and Michael I. Jordan. Is Q-learning provably efficient? In *Neural Information Processing Systems*, 2018.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On bayesian upper confidence bounds for bandit problems. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 2012. PMLR.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 2021.

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, 2019.

Jean-Bernard Lasserre. Simple formula for integration of polynomials on a simplex. *BIT Numerical Mathematics*, 61, 2020.

Pierre Ménard, Omar Darwiche Domingues, Xuedong Shang, and Michal Valko. UCB momentum q-learning: Correcting the bias without forgetting. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7609–7618. PMLR, 2021.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. In *NIPS Deep Learning Workshop*. 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

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

Ian Osband and Benjamin Van Roy. Gaussian-dirichlet posterior dominance in sequential learning. *CoRR*, abs/1702.04126, 2017.

Ian Osband and Benjamin Van Roy. Bootstrapped thompson sampling and deep exploration. *CoRR*, abs/1507.00300, 2015.

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, 2017.

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.

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.

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, 2016b. PMLR.

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.

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.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2721–2730. JMLR. org, 2017.

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, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. 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 2778–2787. PMLR, 2017.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, New York, NY, 1994.

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

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, 2020.

Donald B Rubin. The bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

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.

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.

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

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

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.

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.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. #exploration: A study of count-based exploration for deep reinforcement learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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, 1933.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Chris J. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Zhihan Xiong, Ruoqi Shen, Qiwen Cui, and Simon S. Du. Near-optimal randomized exploration for tabular mdp, 2021.

Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, and Peng Liu. Exploration in deep reinforcement learning: A comprehensive survey. *CoRR*, abs/2109.06668, 2021.

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)*, 2019a.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019b.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

Appendix

Table of Contents

A. Notation

Table 2. Table of notation use throughout the paper

Notation	Meaning
$\mathcal{S}$	state space of size S
$\mathcal{A}$	action space of size A
H	length of one episode
T	number of episodes
B	number of bootstrap 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 fake samples
s_0	fake state
r_0	reward of fake transition
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 bound on the optimal Q-value
$\overline{V}_h^t(s,a)$	upper bound 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

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

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

B. Bayes-UCBVI Proofs

B.1. Concentration events

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

$$\mathcal{E}^\star(\delta) \triangleq \left\{ \forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A}: \quad \mathcal{K}_{\inf}(\widehat{p}_h^t(s,a), p_h V^\star_{h+1}(s,a), V^\star_{h+1}) \le \frac{\beta^\star(\delta, n_h^t(s,a))}{n_h^t(s,a)} \right\},$$

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

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

$$\left. |(\widehat{p}_h^t - p_h) V^\star_{h+1}(s,a)| \le \sqrt{2 \mathrm{Var}_{p_h}(V^\star_{h+1})(s,a) \frac{\beta(\delta, n_h^t(s,a))}{n_h^t(s,a)}} + 3H \frac{\beta(\delta, n_h^t(s,a))}{n_h^t(s,a)} \right\},$$

$$\mathcal{E}^{\mathrm{Var}}(\delta) \triangleq \left\{ \forall t \in \mathbb{N}: \quad \sum_{\ell=1}^{t} \sum_{h=1}^{H} \mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}(s_h^\ell, a_h^\ell)] \le H^2 t + \sqrt{2 H^5 t \beta^{\mathrm{Var}}(\delta, t)} + 3 H^3 \beta^{\mathrm{Var}}(\delta, t) \right\},$$

$$\mathcal{E}(\delta) \triangleq \left\{ \sum_{t=1}^{T} \sum_{h=1}^{H} \left| p_h [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_h^t, a_h^t) - [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_{h+1}^t) \right| \le 2 r_0 H \sqrt{2 H T \beta(\delta)}, \right.$$

$$\left. \sum_{t=1}^{T} \sum_{h=1}^{H} (1 - 1/H)^{H-h+1} \left| p_h [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_h^t, a_h^t) - [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_{h+1}^t) \right| \le 2 \mathrm{e} r_0 H \sqrt{2 H T \beta(\delta)}, \right\}.$$

We also introduce the intersection of these events, $\mathcal{G}(\delta) \triangleq \mathcal{E}^\star(\delta) \cap \mathcal{E}^{\mathrm{KL}}(\delta) \cap \mathcal{E}^{\mathrm{conc}}(\delta) \cap \mathcal{E}^{\mathrm{Var}}(\delta) \cap \mathcal{E}(\delta)$. 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,*

$$\begin{aligned}
\beta^\star(\delta, n) &\triangleq \log(5SAH/\delta) + 3\log(\mathrm{e}\pi(2n+1)), \\
\beta^{\mathrm{KL}}(\delta, n) &\triangleq \log(5SAH/\delta) + \log(\mathrm{e}(1+n)), \\
\beta^{\mathrm{conc}}(\delta, n) &\triangleq \log(5SAH/\delta) + \log(4\mathrm{e}(2n+1)), \\
\beta(\delta) &\triangleq \log(20/\delta), \\
\beta^{\mathrm{Var}}(\delta, t) &\triangleq \log(20\mathrm{e}(2t+1)/\delta),
\end{aligned}$$

it holds that

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}^\star(\delta)] \ge 1 - \delta/5, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \ge 1 - \delta/5, \qquad \mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \ge 1 - \delta/5, \\
&\mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \ge 1 - \delta/5, \qquad \mathbb{P}[\mathcal{E}(\delta)] \ge 1 - \delta/5.
\end{aligned}$$

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

Proof. It follows from Theorem C.4 that $\mathbb{P}[\mathcal{E}^\star(\delta)] \ge 1 - \delta/5$. Applying Theorem C.1 and the union bound over $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$ we get $\mathbb{P}[\mathcal{E}^{\mathrm{KL}}(\delta)] \ge 1 - \delta/5$. Next, Theorem C.6 and the union bound over $h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$ yield $\mathbb{P}[\mathcal{E}^{\mathrm{conc}}(\delta)] \ge 1 - \delta/5$. By Lemma B.2, $\mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \ge 1 - \delta/5$. It remains to estimate $\mathbb{P}[\mathcal{E}(\delta)]$.

Define the following sequences

$$\bar{Z}_{t,h} \triangleq \overline{V}^{t-1}_{h+1}(s_{h+1}^t) - V^*_{h+1}(s_{h+1}^t) - p_h[\overline{V}^{t-1}_{h+1} - V^*_{t+1}](s_h^t, a_h^t), \qquad t \in [T], h \in [H],$$

$$\tilde{Z}_{t,h} \triangleq (1 - 1/H)^{H-h+1} \left(\overline{V}^{t-1}_{h+1}(s_{h+1}^t) - V^*_{h+1}(s_{h+1}^t) - p_h[\overline{V}^{t-1}_{h+1} - V^*_{h+1}](s_h^t, a_h^t) \right), \qquad t \in [T], h \in [H],$$

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

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{Z}_{t,h}\Big| > 2r_0 H\sqrt{2tH\cdot\beta(\delta)}\Big) \le 2\exp(-\beta(\delta)) = \delta/10,$$

$$\mathbb{P}\Big(\Big|\sum_{t=1}^{T}\sum_{h=1}^{H} \bar{Z}_{t,h}\Big| > 2\mathrm{e}r_0 H\sqrt{2tH\cdot\beta(\delta)}\Big) \le 2\exp(-\beta(\delta)) = \delta/10,$$

By the union bound $\mathbb{P}[\mathcal{E}(\delta)] \ge 1 - \delta/5$.

□

Lemma B.2. *Under conditions of Lemma B.1, for any $\delta \in (0,1)$, $\mathbb{P}[\mathcal{E}^{\mathrm{Var}}(\delta)] \ge 1-\delta/5$.*

Proof. For any $\ell \in \mathbb{N}$ define $\mathcal{F}_\ell = \sigma\{(s^j_h, a^j_h), j \le \ell, h \in [H]\}$ and let

$$Y_\ell = \sum_{h=1}^{H} \mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}](s^\ell_h, a^\ell_h) - \sigma V^{\pi_\ell}_1(s^\ell_1),$$

where operator σV is defined in Section E. It is straightforward to check that $(Y_\ell, \mathcal{F}_\ell)_{\ell\in\mathbb{N}}$ is a martingale-difference sequence. Applying Bernstein inequality (Theorem C.6) we get that with probability at least $1-\delta/5$ for any $t \in \mathbb{N}$

$$\left|\sum_{\ell=1}^{t} Y_\ell\right| \le \sqrt{2\sum_{\ell=1}^{t}\mathbb{E}[Y_\ell^2|\mathcal{F}_{\ell-1}]\log(20e(2t+1)/\delta)} + 3H^3\log(20e(2t+1)/\delta).$$

Next we estimate $\mathbb{E}[Y_\ell^2|\mathcal{F}_{\ell-1}]$ in the following way

$$\mathbb{E}[Y_\ell^2|\mathcal{F}_{\ell-1}] \le \mathbb{E}\left[\left(\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}](s^\ell_h, a^\ell_h)\right)^2\middle|\mathcal{F}_{\ell-1}\right] \le H^3\mathbb{E}_{\pi_\ell}\left[\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}](s_h, a_h)\right].$$

By Lemma E.1

$$\mathbb{E}_{\pi_\ell}\left[\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}](s^\ell_h, a^\ell_h)\right] = \mathbb{E}_{\pi_\ell}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h) - V^{\pi_\ell}_1(s_1)\right)^2\right] \le \mathbb{E}_{\pi_\ell}\left[\left(\sum_{h=1}^{H} r_h(s_h,a_h)\right)^2\right] \le H^2,$$

Since $\beta^{\mathrm{Var}}(\delta,t) = \log(20e(2t+1)/\delta)$, we have

$$\sum_{\ell=1}^{t} Y_\ell \le \sqrt{2H^5 t\beta^{\mathrm{Var}}(\delta,t)} + 3H^3\beta^{\mathrm{Var}}(\delta,t).$$

Finally, by Lemma E.1

$$\sum_{\ell=1}^{t}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V^{\pi_\ell}_{h+1}](s^\ell_h, a^\ell_h) = \sum_{\ell=1}^{t} Y_\ell + \sum_{\ell=1}^{t}\sigma V^{\pi_\ell}_1(s^\ell_1) \le \sqrt{2H^5 t\beta^{\mathrm{Var}}(\delta,t)} + 3H^3\beta^{\mathrm{Var}}(\delta,t) + H^2 t.$$

□

Lemma B.3. *Assume conditions of Lemma B.1. Then conditioned on event $\mathcal{E}^{\mathrm{KL}}(\delta)$, for any $f\colon \mathcal{S} \to [0, r_0 H]$, $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$(\widehat{p}^t_h - p_h)f(s,a) \le \frac{1}{H}p_h f(s,a) + \frac{5r_0 H^2 S\cdot\beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)},$$

$$\|\widehat{p}^t_h(s,a) - p_h(s,a)\|_1 \le \sqrt{\frac{2S\cdot\beta^{\mathrm{KL}}(\delta, n^t_h(s,a))}{n^t_h(s,a)}}.$$

Proof. We apply Lemma E.2 and Lemma E.3 to obtain

$$
\begin{aligned}
(\widehat{p}_h^t - p_h) f(s,a) &\le \sqrt{2\mathrm{Var}_{\widehat{p}_h^t}[f](s,a)\cdot \mathrm{KL}(\widehat{p}_h^t, p_h)} + \frac{2Hr_0}{3}\,\mathrm{KL}(\widehat{p}_h^t, p_h) \\
&\le 2\sqrt{\mathrm{Var}_{p_h}[f](s,a)\cdot \mathrm{KL}(\widehat{p}_h^t, p_h)} + \left(2\sqrt{2}+\frac{2}{3}\right) H r_0\,\mathrm{KL}(\widehat{p}_h^t, p_h).
\end{aligned}
$$

Since $0 \le f(s) \le r_0 H$ we get

$$
\mathrm{Var}_{p_h}[f](s,a) \le p_h[f^2](s,a) \le r_0 H \cdot p_h f(s,a).
$$

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

$$
(\widehat{p}_h^t - p_h) f(s,a) \le \frac{1}{H} p_h f(s,a) + (H^2 + 2\sqrt{2} r_0 H + 2r_0 H/3)\,\mathrm{KL}(\widehat{p}_h^t, p_h) \le \frac{1}{H} p_h f(s,a) + 5 r_0 H^2\,\mathrm{KL}(\widehat{p}_h^t, p_h).
$$

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 conditioned on the event $\mathcal{E}^\star(\delta)$ our estimate of Q-function $\overline{Q}_h^t(s,a)$ is optimistic that is $\overline{Q}_h^t(s,a) \ge Q_h^\star(s,a)$ for any $t \le T, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$.

For any $\beta > 0, p \in \Delta_{S'-1}$ and $f : \mathcal{S}' \to \mathbb{R}$ define

$$
U^{\mathcal{K}_{\inf}}(\beta, p, f) = \sup\{\mu \ge pf : \mathcal{K}_{\inf}(p, \mu, f) \le \beta\}.
$$

First we are going to prove that $U^{\mathcal{K}_{\inf}}(\beta^\star(\delta, n_h^t(s,a))/\overline{n}_h^t(s,a), \overline{p}_h^t(s,a), V_{h+1}^\star)$ defines an upper confidence bound for $p_h V_{h+1}^\star(s,a)$.

Lemma B.4. *Conditioned on the event $\mathcal{E}^\star(\delta)$, for any $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S}\times\mathcal{A}$,*

$$
p_h V_{h+1}^\star(s,a) \le U^{\mathcal{K}_{\inf}}\left(\frac{\beta^\star(\delta, n_h^t(s,a))}{\overline{n}_h^t(s,a)}, \overline{p}_h^t(s,a), V_{h+1}^\star\right),
$$

where event $\mathcal{E}^\star(\delta)$ and function $\beta^\star(\delta, n)$ were defined in Lemma B.1.

Proof. By Lemma C.2 we have for any $\overline{p}_h^t V_h^\star(s,a) < u < r_0(H-h)$

$$
\begin{aligned}
\overline{n}_h^t(s,a)\,\mathcal{K}_{\inf}(\overline{p}_h^t(s,a), u, V_{h+1}^\star) &= \overline{n}_h^t(s,a) \max_{\lambda\in[0,1]} \mathbb{E}_{s'\sim\overline{p}_h^t(s,a)}\left[\log\left(1-\lambda\frac{V_{h+1}^\star(s')-u}{r_0(H-h)-u}\right)\right] \\
&\le \max_{\lambda\in[0,1]} n_0\log(1-\lambda) + (\overline{n}_h^t(s,a)-n_0)\max_{\lambda\in[0,1]}\mathbb{E}_{s'\sim\widehat{p}_h^t(s,a)}\left[\log\left(1-\lambda\frac{V_{h+1}^\star(s')-u}{r_0(H-h)-u}\right)\right] \\
&\le (\overline{n}_h^t(s,a)-n_0)\max_{\lambda\in[0,1]}\mathbb{E}_{s'\sim\widehat{p}_h^t(s,a)}\left[\log\left(1-\lambda\frac{V_{h+1}^\star(s')-u}{H-h-u}\right)\right] \\
&= (\overline{n}_h^t(s,a)-n_0)\,\mathcal{K}_{\inf}(\widehat{p}_h^t(s,a), u, V_{h+1}^\star) = n_h^t(s,a)\,\mathcal{K}_{\inf}(\widehat{p}_h^t(s,a), u, V_{h+1}^\star).
\end{aligned}
$$

By the definition of event $\mathcal{E}^\star(\delta)$ we have for any $t\in\mathbb{N}, h\in[H], (s,a)\in\mathcal{S}\times\mathcal{A}$,

$$
n_h^t(s,a)\,\mathcal{K}_{\inf}(\widehat{p}_h^t(s,a), p_h V_{h+1}^\star(s,a), V_{h+1}^\star) \le \beta^\star(\delta, n_h^t(s,a)),
$$

hence $\overline{n}_h^t(s,a)\,\mathcal{K}_{\inf}(\overline{p}_h^t(s,a), p_h V_{h+1}^\star(s,a), V_{h+1}^\star) \le \beta^\star(\delta, n_h^t(s,a))$. Therefore a value $p_h V_{h+1}^\star(s,a)$ is feasible for optimization problem in the definition of $U^{\mathcal{K}_{\inf}}$. □

For the further results we have to guarantee that a number of observations of the fake state s_0 is large enough to apply anti-concentration result of Dirichlet distribution. Define constant

$$
c_{n_0} = \frac{1}{(\sqrt{2\pi}-1)^2}\cdot\left(\frac{2\sqrt{2}}{\sqrt{\log(17/16)}} + \frac{98\sqrt{6}}{9}\right)^2 + \frac{\log(10\pi)}{\log(17/16)}. \tag{4}
$$

Lemma B.5. *Let $n_0 \ge c_{n_0} + \log_{17/16}(T)$, where c_{n_0} is defined in (4), and $r_0 \ge 2$, and assume conditions of Lemma B.1. Then on the event $\mathcal{E}^\star(\delta)$, it holds for any $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$p_h V^\star_{h+1}(s,a) \le \mathbb{Q}_{p \sim \rho^t_h(s,a)}(pV^\star_{h+1}(s,a), \kappa^t_h(s,a)),$$

where $\kappa^t_h(s,a) = 1 - \frac{C_\kappa \delta}{SAH[2n^t_h(s,a)+1]^3[\overline{n}^t_h(s,a)]^{3/2}}$ with an absolute constant $C_\kappa = 1/(5 \cdot (\mathrm{e}\pi)^3)$.

Proof. To simplify notations we set $\overline{n} = \overline{n}^t_h(s,a)$ and $n = n^t_h(s,a)$. Note that $\rho^t_h(s,a)$ is a Dirichlet distribution $\mathcal{D}\mathrm{ir}(\{\overline{n}^t_h(s'|s,a)\}_{s' \in \mathcal{S}'})$. Since $\overline{n}^t_h(s_0|s,a) = n_0 \ge c_{n_0} + \log_{17/16}(T)$, we may apply Theorem D.2 if $\overline{n} \ge 2n_0$: for any $\overline{p}^t_h V^\star_{h+1}(s,a) \le u < r_0(H-h)$

$$\mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge u\right) \ge \exp(-\overline{n}\,\mathcal{K}_{\mathrm{inf}}(\overline{p}^t_h(s,a), u, V^\star_{h+1}) - 3/2 \log \overline{n}). \tag{5}$$

Notice that the same inequality also holds for $u < \overline{p}^t_h V^\star_{h+1}(s,a)$ because $\mathcal{K}_{\mathrm{inf}}(\overline{p}^t_h(s,a), u, V^\star_{h+1}) = 0$ and

$$\mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge u\right) \ge \mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge \overline{p}^t_h V^\star_{h+1}(s,a)\right).$$

Let $u' = U^{\mathcal{K}_{\mathrm{inf}}}(\beta^\star(\delta,n)/\overline{n}, \overline{p}^t_h(s,a), V^\star_{h+1})$. Fix arbitrary $\varepsilon > 0$ and set $u = u' - \varepsilon$. This choice implies that $\overline{n}\,\mathcal{K}_{\mathrm{inf}}(\overline{p}^t_h(s,a), u, V^\star_{h+1}) \le \beta^\star(\delta,n)$, and together with (5) yields

$$\mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge u\right) \ge \exp(-\beta^\star(\delta,n) - 3/2\log(\overline{n})) \ge \frac{C_\kappa \delta}{SAH[2n^t_h(s,a)+1]^3[\overline{n}^t_h(s,a)]^{3/2}}.$$

By Lemma E.5 and definition of $\kappa^t_h(s,a)$, $\mathbb{Q}_{p \sim \rho^t_h(s,a)}(pV^\star_{h+1}(s,a), \kappa^t_h(s,a)) \ge u' - \varepsilon$. Allowing $\varepsilon \to 0$ we have $\mathbb{Q}_{p \sim \rho^t_h(s,a)}(pV^\star_{h+1}(s,a), \kappa^t_h(s,a)) \ge u'$. It remains to apply Lemma B.4 to conclude the statement in the case $\overline{n} \ge 2n_0$.

To handle the case $\overline{n} < 2n_0$ we remark that

$$\mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge p_h V^\star_{h+1}(s,a)\right) \ge \mathbb{P}_{\xi \sim B(n_0, \overline{n}-n_0)}(r_0(H-h)\xi \ge H-h) \ge \mathbb{P}_{\xi \sim B(n_0, \overline{n}-n_0)}\left(\xi \ge \frac{1}{2}\right),$$

where we used an upper bound $p_h V^\star_{h+1}(s,a) \le H-h$ and a lower bound $V^\star_{h+1}(s) \ge 0$ for $s \in \mathcal{S}$ and $V^\star_{h+1}(s_0) = r_0(H-h)$. By the result of Groeneveld and Meeden (1977) we have that for $n_0 \le \overline{n} - n_0$ we have that the median m of ξ is greater than the mode $(n_0-1)/(\overline{n}-2)$. Since $2n_0 > \overline{n}$, we have that $m \ge 1/2$, thus

$$\begin{aligned}\mathbb{P}_{p \sim \rho^t_h(s,a)}\left(pV^\star_{h+1} \ge p_h V^\star_{h+1}(s,a)\right) &\ge \mathbb{P}_{\xi \sim B(n_0, \overline{n}-n_0)}\left(\xi \ge \frac{1}{2}\right) \ge \mathbb{P}_{\xi \sim B(n_0, \overline{n}-n_0)}(\xi \ge m) = \frac{1}{2} \\ &\ge \frac{C_\kappa \delta}{SAH[2n^t_h(s,a)+1]^3[\overline{n}^t_h(s,a)]^{3/2}}.\end{aligned}$$

Lemma E.5 concludes the statement. □

Proposition B.6 (Optimism)**.** *Let $n_0 = \lceil c_{n_0} + \log_{17/16}(T) \rceil$, where c_{n_0} is an absolute constant defined in (4). Furthermore, let $r_0 = 2$ and assume that conditions of Lemma B.1 are satisfied. Then $\overline{Q}^t_h(s,a) \ge Q^\star_h(s,a)$ on the event $\mathcal{E}^\star(\delta)$ for any $t \le T, h \in [H]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof. We proceed using backward induction over h. For $h = H+1$, $\overline{Q}^t_h(s,a) = Q^\star_h(s,a) = 0$. Let $h \le H$. Note that

$$\overline{Q}^t_h(s,a) - Q^\star_h(s,a) = \mathbb{Q}_{p \sim \rho^t_h(s,a)}(p\overline{V}^t_{h+1}(s,a), \kappa^t_h(s,a)) - p_h V^\star_{h+1}(s,a). \tag{6}$$

Induction hypothesis implies that

$$V^\star_{h+1}(s) = Q^\star_{h+1}(s, \pi^\star(s)) \le \overline{Q}^t_{h+1}(s, \pi^\star(s)) \le \overline{V}^t_{h+1}(s),$$

and hence

$$\mathbb{Q}_{p \sim \rho^t_h(s,a)}(p\overline{V}^t_{h+1}(s,a), \kappa^t_h(s,a)) \ge \mathbb{Q}_{p \sim \rho^t_h(s,a)}(pV^\star_{h+1}(s,a), \kappa^t_h(s,a)). \tag{7}$$

Equation (6), inequality (7) and Lemma B.5 imply the statement. □

Next we formulate key inequality for the further proof of regret bound.

Corollary B.7. *Let $n_0 = \lceil c_{n_0} + \log_{17/16}(T) \rceil$ and $r_0 = 2$. Under conditions of Lemma B.1, it holds on the event $\mathcal{E}^\star(\delta)$ for any $t \in \mathbb{N}, h \in [H], (s,a) \in \mathcal{S} \times \mathcal{A}$,*

$$
\begin{aligned}
p_h V^\star_{h+1}(s,a) &\le \mathbb{Q}_{p \sim \rho_h^t(s,a)}(p\overline{V}^t_{h+1}(s,a), \kappa_h^t(s,a)) \\
&\le \overline{p}_h^t \overline{V}^t_{h+1}(s,a) + 2\sqrt{\frac{\mathrm{Var}_{\overline{p}_h^t}[\overline{V}^t_{h+1}](s,a)\log\left(\frac{1}{1-\kappa_h^t(s,a)}\right)}{\overline{n}_h^t(s,a)}} + \frac{2r_0 H\sqrt{2}\log\left(\frac{1}{1-\kappa_h^t(s,a)}\right)}{\overline{n}_h^t(s,a)},
\end{aligned}
$$

where $\kappa_h^t(s,a) = 1 - \frac{C_\kappa \delta}{SAH[2n_h^t(s,a)+1]^3[\overline{n}_h^t(s,a)]^{3/2}}$ with an absolute constant $C_\kappa = 1/(5\cdot(\mathrm{e}\pi)^3)$ and c_{n_0} defined in (4).

Proof. The first inequality immediately follows from Proposition B.6. The second inequality follows from Lemma C.8, where we take $\delta = 1 - \kappa_h^t(s,a)$, $f = \overline{V}^t_{h+1}$, and Lemma E.5. □

B.3. Proof of Theorem 3.1

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

$$
L = \max\left\{n_0, \log(TH), \max_{t\in[T], h\in[H]} \log\left(\frac{1}{1-\hat{\kappa}_h^t}\right), \beta^\star(\delta, T), \beta^{\mathrm{KL}}(\delta, T), \beta^{\mathrm{conc}}(\delta, T), \beta(\delta), \beta^{\mathrm{Var}}(\delta, T), 1\right\}. \tag{8}
$$

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

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

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

where

$$
\begin{aligned}
A_h^T &= 2\mathrm{e}\sqrt{L}\sum_{t=1}^T\sum_{h'=h}^H \sqrt{\mathrm{Var}_{\overline{p}_{h'}^{t-1}}[\overline{V}^{t-1}_{h'+1}](s_{h'}^t, a_{h'}^t)\frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}}, \\
B_h^T &= \mathrm{e}\sqrt{2L}\sum_{t=1}^T\sum_{h'=h}^H \sqrt{\mathrm{Var}_{p_{h'}}[V^\star_{h'+1}](s_{h'}^t, a_{h'}^t)\frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t}}, \\
C_h^T &= 21\mathrm{e}H^2 S\cdot L\cdot\sum_{t=1}^T\sum_{h=h'}^H \frac{\mathbb{1}\{N_{h'}^t > 0\}}{N_{h'}^t},
\end{aligned}
$$

and L is defined in (8).

Proof. By the greedy choice of action, formula (1) for $\overline{Q}$ and Bellman's equations

$$
\begin{aligned}
\delta_h^t &= r_h(s_h^t, a_h^t) + \mathbb{Q}_{p\sim\rho_h^{t-1}(s_h^t, a_h^t)}(p\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t), \kappa_h^{t-1}(s_h^t, a_h^t)) - r_h(s_h^t, a_h^t) - p_h V^{\pi_t}_{h+1}(s_h^t, a_h^t) \\
&= \mathbb{Q}_{p\sim\rho_h^{t-1}(s_h^t, a_h^t)}(p\overline{V}^{t-1}_{h+1}(s,a), \kappa_h^{t-1}(s_h^t, a_h^t)) - (\overline{p}_h^{t-1} - \overline{p}_h^{t-1})\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t) \\
&- (\widehat{p}_h^{t-1} - \widehat{p}_h^{t-1})\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t) - p_h\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t) + p_h[\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_h^t, a_h^t) \\
&= \underbrace{\mathbb{Q}_{p\sim\rho_h^{t-1}(s_h^t, a_h^t)}(p\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t), \kappa_h^{t-1}(s_h^t, a_h^t)) - \overline{p}_h^{t-1}\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t)}_{\textbf{(A)}} + \underbrace{[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}^{t-1}_{h+1}(s_h^t, a_h^t)}_{\textbf{(B)}} \\
&+ \underbrace{(\widehat{p}_h^{t-1} - p_h)[\overline{V}^{t-1}_{h+1} - V^\star_{h+1}](s_h^t, a_h^t)}_{\textbf{(C)}} + \underbrace{(\widehat{p}_h^{t-1} - p_h)V^\star_{h+1}(s_h^t, a_h^t)}_{\textbf{(D)}} \\
&+ \underbrace{p_h[\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_h^t, a_h^t) - [\overline{V}^{t-1}_{h+1} - V^{\pi_t}_{h+1}](s_{h+1}^t)}_{\xi_h^t} + \delta_{h+1}^t.
\end{aligned}
$$

It is easy to see that ξ_h^t appears in the definition of event $\mathcal{E}(\delta) \subseteq \mathcal{G}(\delta)$.

We analyse each term in this representation under assumption $N_h^t > 0$.

Term $(\mathbf{A})$**.** Then to estimate this term we apply the second inequality in Corollary B.7. We obtain

$$\mathbb{Q}_{p\sim\rho_h^{t-1}(s_h^t,a_h^t)}(p\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t),\hat{\kappa}_h^t) - \overline{p}_h^{t-1}\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t) \le 2\sqrt{\frac{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t,a_h^t)\log\left(\frac{1}{1-\hat{\kappa}_h^t}\right)}{\overline{N}_h^t}} + \frac{2r_0\sqrt{2}H\log\left(\frac{1}{1-\hat{\kappa}_h^t}\right)}{\overline{N}_h^t}.$$

Note that this term acts very similar to Bernstein-type bonuses in `UCBVI` algorithm.

Term $(\mathbf{B})$**.** The bound follows directly from the definition of $\overline{p}_h^t$ and $\widehat{p}_h^t$. Indeed,

$$[\overline{p}_h^{t-1} - \widehat{p}_h^{t-1}]\overline{V}_{h+1}^{t-1}(s_h^t,a_h^t) = \frac{n_0}{\overline{N}_h^t}\cdot(r_0H) + \sum_{s'\in\mathcal{S}}\left(\frac{N_h^t(s')}{\overline{N}_h^t} - \frac{N_h^t(s')}{N_h^t}\right)\cdot\overline{V}_{h+1}^t(s') \le \frac{r_0HL}{\overline{N}_h^t}.$$

Term $(\mathbf{C})$**.** To estimate this term we first note that by Proposition B.6, $\overline{V}_{h+1}^{t-1}(s) - V_{h+1}^\star(s) \ge 0$ for any $s\in\mathcal{S}$. Hence, we may use Lemma B.3 with $f = \overline{V}_{h+1}^t - V_{h+1}^\star$. We obtain

$$\begin{aligned}(\widehat{p}_h^{t-1} - p_h)[\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t,a_h^t) &\le \frac{1}{H}p_h[\overline{V}_{h+1}^{t-1} - V_{h+1}^\star](s_h^t,a_h^t) + \frac{5r_0H^2S\cdot\beta^{\mathrm{KL}}(\delta,N_h^t)}{N_h^t}\\ &\le \frac{1}{H}(\xi_h^t + \delta_h^t) + \frac{5r_0H^2S\cdot L}{N_h^t}.\end{aligned}$$

Term $(\mathbf{D})$**.** By the definition of event $\mathcal{E}^{\mathrm{conc}}(\delta) \subseteq \mathcal{G}(\delta)$

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

Collecting bounds for the terms $(\mathbf{A})$–$(\mathbf{D})$ we get

$$\begin{aligned}\delta_h^t \le \left(1+\frac{1}{H}\right)\delta_{h+1}^t + \left(1+\frac{1}{H}\right)\xi_h^t + 2\sqrt{\mathrm{Var}_{\overline{p}_h^t}[\overline{V}_{h+1}^t](s_h^t,a_h^t)\frac{L}{\overline{N}_h^t}}\\ + \sqrt{2\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t)\frac{L}{N_h^t}} + \frac{(2r_0\sqrt{2}H + r_0H + 5r_0H^2S + 3H)L}{N_h^t}.\end{aligned}$$

Notice that in the case $N_h^t = 0$ we have a trivial bound $\delta_h^t \le r_0H$. However, this case might appear at most SAH times in the summation and thus we can handle this case by additive r_0SAH^2 error term.

Define $\gamma_h = (1+1/H)^{H-h+1}$. Notice that $\gamma_h < \mathrm{e}$, $1/\overline{N}_h^t < 1/N_h^t$, $r_0 = 2$, $H \le H^2S$. After summation, we have

$$\begin{aligned}\overline{\mathfrak{R}}_h^T \le& \sum_{t=1}^T\sum_{h'=h}^H \gamma_{h'}\xi_{h'}^t + r_0H^2SA\\ &+ 2\mathrm{e}\sqrt{L}\sum_{t=1}^T\sum_{h'=h}^H\sqrt{\mathrm{Var}_{\overline{p}_{h'}^{t-1}}[\overline{V}_{h'+1}^{t-1}](s_{h'}^t,a_{h'}^t)\frac{\mathbb{1}\{N_{h'}^t>0\}}{N_{h'}^t}} && \triangleq A_h^T\\ &+ \mathrm{e}\sqrt{2L}\sum_{t=1}^T\sum_{h'=h}^H\sqrt{\mathrm{Var}_{p_{h'}}[V_{h'+1}^\star](s_{h'}^t,a_{h'}^t)\frac{\mathbb{1}\{N_{h'}^t>0\}}{N_{h'}^t}} && \triangleq B_h^T\\ &+ 21\mathrm{e}H^2S\cdot L\cdot\sum_{t=1}^T\sum_{h=h'}^H\frac{\mathbb{1}\{N_{h'}^t>0\}}{N_{h'}^t}. && \triangleq C_h^T\end{aligned}$$

Finally, by definition of the event $\mathcal{E}(\delta)$ we get

$$\sum_{t=1}^{T}\sum_{h'=h}^{H}\gamma_{h'}\xi_{h'}^t \le 4\mathrm{e}\cdot H\sqrt{2HTL}.$$

$\square$

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 to $n_h^{t-1}(s,a)$.

For the first sum we use a bound on harmonic numbers

$$\sum_{t=1}^{T}\frac{\mathbb{1}\{n_h^{t-1}(s_h^t,a_h^t)>0\}}{n_h^{t-1}(s_h^t,a_h^t)} = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\sum_{n=1}^{n_h^{t-1}(s,a)}\frac{1}{n} \le SA(\log(T)+1) \le 2SAL.$$

To finish the proof of the first inequality it remains to take a sum w.r.t h. For the second sum we use the following integral bound

$$\sum_{t=1}^{T}\frac{\mathbb{1}\{n_h^{t-1}(s_h^t,a_h^t)>0\}}{\sqrt{n_h^{t-1}(s_h^t,a_h^t))}} = \sum_{(s,a)\in\mathcal{S}}\sum_{n=1}^{n_h^{t-1}(s,a)}\frac{1}{\sqrt{n}} \le \sum_{(s,a)\in\mathcal{S}}2\sqrt{n_h^{t-1}(s,a)+1}. \tag{9}$$

Since $\sum_{s,a} n_h^{t-1}(s,a) = t-1$, the last sum is maximized if $n_h^{t-1}(s,a) = (t-1)/(SA)$. This implies the second statement. $\square$

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

$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t,a_h^t)\mathbb{1}\{N_h^t>0\} \le 2H^2T + 2H^2U_1^T + 22H^3S^2AL^2 + 32H^3S\sqrt{2ATL},$$
$$\sum_{t=1}^{T}\sum_{h=1}^{H}\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t) \le 2H^2T + 2H^2U_1^T + 6H^3L + 8\sqrt{2H^5TL}.$$

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

Proof. We apply the second inequality in Lemma E.4,

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

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

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

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

$$(\mathbf{X}) \le 2r_0^2 H^2 \sum_{t=1}^{T}\sum_{h=1}^{H} \|\overline{p}_h^t(s_h^t,a_h^t) - p_h(s_h^t,a_h^t)\|_1 \le 16H^3S^2AL^2 + 24H^3S\sqrt{2ATL}.$$

Next, we bound $(\mathbf{W})$ using the first inequality in Lemma E.4. We get

$$(\mathbf{W}) \le 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[V_{h+1}^{\pi_t}](s_h^t,a_h^t)}_{(\mathbf{Y})} + 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} r_0 H p_h\left|\overline{V}_{h+1}^{t-1} - V_{h+1}^{\pi_t}\right|(s_h^t,a_h^t)}_{(\mathbf{Z})}.$$

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

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

By Proposition B.6 we have $\overline{V}_{h+1}^t(s) \ge V_{h+1}^{\pi_t}(s)$ for any $s \in \mathcal{S}$. By the definition of ξ_h^t, δ_h^t and definition of event $\mathcal{E}$ term $(\mathbf{Z})$ could be bounded as follows

$$\begin{aligned}(\mathbf{Z}) &\le \sum_{t=1}^{T}\sum_{h=1}^{H} 2H(\xi_h^t + \delta_{h+1}^t) \\ &\le 2r_0H^2\sqrt{2TL} + 2H\sum_{h=1}^{H}\overline{\mathfrak{R}}_{h+1}^T \le 4H^2\sqrt{2TL} + 2H^2U_1^T.\end{aligned}$$

Here the last inequality follows from Lemma B.8. Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t,a_h^t)\mathbb{1}\{N_h^t > 0\} &\le (\mathbf{X}) + 2\cdot(\mathbf{Y}) + 2\cdot(\mathbf{Z}) \\ &\le 2H^2T + 2H^2U_1^T + 22H^3S^2AL^2 + (24+8)H^3S\sqrt{2ATL} \\ &\le 2H^2T + 2H^2U_1^T + 22H^3S^2AL^2 + 32H^3S\sqrt{2ATL}.\end{aligned}$$

To bound the second inequality one may apply the first inequality in Lemma E.4. We get

$$\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t) \le 2\underbrace{\sum_{t=1}^{T}\sum_{h=1}^{H} \mathrm{Var}_{p_h}[V_{h+1}^{\pi_t}](s_h^t,a_h^t)}_{(\mathbf{Y})} + 2\sum_{t=1}^{T}\sum_{h=1}^{H} r_0 H p_h\left|V_{h+1}^\star - V_{h+1}^{\pi_t}\right|(s_h^t,a_h^t).$$

Note that by Proposition B.6 the second term is bounded by $(\mathbf{Z})$. Thus

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

$\square$

Lemma B.11. *Under conditions of Lemma B.8, it holds on the event $\mathcal{G}(\delta)$,*

$$
\begin{aligned}
A_1^T &\le 4\mathrm{e}\sqrt{H^3SAT}\cdot L + 4\mathrm{e}\sqrt{H^3SAU_1^T}\cdot L + 14\mathrm{e}H^2S^{3/2}AL^2 + 20\mathrm{e}H^2SA^{3/4}T^{1/4}L^{5/4},\\
B_1^T &\le 4\mathrm{e}\sqrt{H^3SAT}\cdot L + 4\mathrm{e}\sqrt{H^3SAU_1^T}\cdot L + 8\mathrm{e}H^2S^{1/2}A^{1/2}L^2 + 10\mathrm{e}H^{7/4}S^{1/2}A^{1/2}T^{1/4}L^{5/4},\\
C_1^T &\le 42\mathrm{e}H^3S^2AL^2 = \widetilde{\mathcal{O}}(H^3S^2A).
\end{aligned}
$$

Proof. To bound A_1^T we apply the Cauchy—Schwartz inequality, Lemma B.10, Lemma B.9 and inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}, a,b\ge 0$,

$$
\begin{aligned}
\sum_{t=1}^T\sum_{h=1}^H\sqrt{\mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t,a_h^t)\frac{\mathbb{1}\{N_h^t>0\}}{N_h^t}} &\le \sqrt{\sum_{t=1}^T\sum_{h=1}^H \mathrm{Var}_{\overline{p}_h^{t-1}}[\overline{V}_{h+1}^{t-1}](s_h^t,a_h^t)\mathbb{1}\{N_h^t>0\}}\cdot\sqrt{\sum_{t=1}^T\sum_{h=1}^H\frac{\mathbb{1}\{N_h^t>0\}}{N_h^t}}\\
&\le \sqrt{2H^2T + 2H^2U_1^T + 22H^3S^2AL^2 + 32H^3S\sqrt{2ATL}}\cdot\sqrt{2SAHL}\\
&\le 2\sqrt{H^3SATL} + 2\sqrt{H^3SAU_1^TL} + 7H^2S^{3/2}AL^{3/2} + 10H^2SA^{3/4}T^{1/4}L^{3/4}.
\end{aligned}
$$

Similarly, the term B_1^T may be estimated as follows

$$
\begin{aligned}
\sum_{t=1}^T\sum_{h=1}^H\sqrt{\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t)\frac{\mathbb{1}\{N_h^t>0\}}{N_h^t}} &\le \sqrt{\sum_{t=1}^T\sum_{h=1}^H\mathrm{Var}_{p_h}[V_{h+1}^\star](s_h^t,a_h^t)}\cdot\sqrt{\sum_{t=1}^T\sum_{h=1}^H\frac{\mathbb{1}\{N_h^t>0\}}{N_h^t}}\\
&\le \sqrt{2H^2T + 2H^2U_1^T + 8\sqrt{2H^5TL} + 6H^3L}\cdot\sqrt{2SAH\cdot L}\\
&\le 2\sqrt{H^3SATL} + 2\sqrt{H^3SAU_1^TL} + 4H^2L\sqrt{SA} + 5H^{7/4}T^{1/4}L^{3/4}\sqrt{SA}.
\end{aligned}
$$

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

$$
C_1^T \le 21\mathrm{e}H^2S\cdot L\cdot 2SAHL \le 42\mathrm{e}H^3S^2AL^2.
$$

□

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

By Proposition B.6 and Lemma B.8

$$
\mathfrak{R}^T = \sum_{t=1}^T V_h^\star(s_1^t) - V_h^{\pi_t}(s_1^t) \le \sum_{t=1}^T \overline{V}_h^{t-1}(s_1^t) - V_h^{\pi_t}(s_1^t) = \overline{\mathfrak{R}}_1^T \le U_1^T = A_1^T + B_1^T + C_1^T + 4\mathrm{e}\sqrt{2H^3TL} + 2\mathrm{e}SAH^2. \quad (10)
$$

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

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

Furthermore,

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

We obtain the following bounds

$$
\begin{aligned}
A_1^T &\le 24\mathrm{e}\sqrt{H^3SAT}\cdot L + 4\mathrm{e}\sqrt{H^3SAU_1^T}\cdot L + 14\mathrm{e}H^3S^2AL^2,\\
B_1^T &\le 14\mathrm{e}\sqrt{H^3SAT}\cdot L + 4\mathrm{e}\sqrt{H^3SAU_1^T}\cdot L + 8\mathrm{e}H^3S^2AL^2,\\
C_1^T &\le 42\mathrm{e}H^3S^2AL^2 \le 42\mathrm{e}H^3S^2AL^2.
\end{aligned}
$$

Hence, by a bound $SAH^2 \leq H^3 S^2 AL^2$

$$U_1^T \leq 38\mathrm{e}\sqrt{H^3 SAT} \cdot L + 8\mathrm{e}\sqrt{H^3 SAU_1^T} \cdot L + 66\mathrm{e}H^3 S^2 AL^2 + 4\mathrm{e}\sqrt{2} \cdot \sqrt{H^3 TL}.$$

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

$$U_1^T \leq 176\mathrm{e}\sqrt{H^3 SAT} \cdot L + 264\mathrm{e}H^3 S^2 AL^2 + 256\mathrm{e}^2 H^3 SAL^2.$$

The last inequality and (10) imply that

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

□

C. Deviation Inequalities

C.1. Deviation inequality for categorical distributions

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

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

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 $\delta \in [0,1]$,*

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

C.2. Deviation inequality for categorical weighted sum

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

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

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

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

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

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

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

We are now ready to state the deviation inequality for the $\mathcal{K}_{\inf}$ 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 $\delta \in [0,1]$,*

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

Proof. First if $pf = b$ then $f(k) = b$ for all k such that $p_k > 0$. In this case $\mathcal{K}_{\inf}(\widehat{p}_n, pf, f) = 0$ for all n and the result is trivially true. We thus assume now that $pf < b$.

The proof is a combination of the one of Proposition 13 by Garivier et al. (2018) and the method of mixtures. We first define the martingale

$$M_n^\lambda = \exp\left(\sum_{\ell=1}^{n}\log\left(1-\lambda\frac{f(X_\ell)-pf}{b-pf}\right)\right),$$

with the convention $M_0^\lambda = 1$. Indeed if we denote by $\mathcal{F}_n = \sigma(X_1,\ldots,X_n)$ the information available at time n, we have

$$\mathbb{E}\big[M_n^\lambda | \mathcal{F}_{n-1}\big] = \mathbb{E}\left[1-\lambda\frac{f(X_n)-pf}{b-pf}\right]M_{n-1}^\lambda = M_{n-1}^\lambda\,.$$

We fix a real number $\gamma_j = 1/(2j)$ for $j \in \mathbb{N}^*$and let S_j be the set

$$S_j = \left\{ \frac{1}{2} - \left\lfloor \frac{1}{2\gamma_j} \right\rfloor \gamma_j, \dots, \frac{1}{2} - \gamma_j, \, \frac{1}{2}, \, \frac{1}{2} + \gamma_j, \dots, \frac{1}{2} + \left\lfloor \frac{1}{2\gamma_j} \right\rfloor \gamma_j \right\}.$$

The cardinality of this set S_j is bounded by $1 + 2j$. We choose a prior on λ the mixture of uniform distribution over this grid: $6/\pi^2 \sum_{j=1}^{\infty} 1/j^2 \mathcal{U}(S_j)$. Thus we consider the integrated martingale

$$\begin{aligned} M_n &= \frac{6}{\pi^2} \sum_{j=1}^{\infty} \frac{1}{j^2} \sum_{\lambda \in S_j} \frac{1}{|S_j|} M_n^\lambda \\ &\geq \frac{6}{\pi^2 n^2 |S_n|} \max_{\lambda \in S_n} M_n^\lambda \\ &\geq \frac{6}{\pi^2 (1+2n)^3} \max_{\lambda \in S_n} M_n^\lambda \,. \end{aligned} \tag{11}$$

Lemma C.5 below indicates that for all $\lambda \in [0, 1]$, there exists a $\lambda' \in S_n$ such that for all $x \in [0, b]$,

$$\log\left(1 - \lambda \frac{x - pf}{b - pf}\right) \leq 2\gamma_n + \log\left(1 - \lambda' \frac{x - pf}{b - pf}\right). \tag{12}$$

Now, a combination of the variational formula of Lemma C.2 and of the inequality (12) yields a finite maximum as an upper bound on $\mathcal{K}_{\text{inf}}(\widehat{p}_n, pf, f)$

$$\begin{aligned} \mathcal{K}_{\text{inf}}(\widehat{p}_n, pf, f) &= \max_{0 \leq \lambda \leq 1} \frac{1}{n} \sum_{\ell=1}^{n} \log\left(1 - \lambda \frac{X_\ell - pf}{b - pf}\right) \\ &\leq 2\gamma_n + \max_{\lambda' \in S_n} \frac{1}{n} \sum_{k=1}^{n} \log\left(1 - \lambda' \frac{X_\ell - pf}{b - pf}\right). \end{aligned}$$

Thanks to the definition of the martingale M_n^λ we obtain

$$\max_{\lambda \in S_n} M_n^\lambda \geq e^{-2n\gamma_n} e^{n\, \mathcal{K}_{\text{inf}}(\widehat{p}_n, pf, f)} = e^{-1} e^{n\, \mathcal{K}_{\text{inf}}(\widehat{p}_n, pf, f)} \,.$$

Combining this inequality with (11) yields

$$M_n \geq \frac{6}{e\pi^2 (1+2n)^3} e^{n\, \mathcal{K}_{\text{inf}}(\widehat{p}_n, pf, f)} \,.$$

Since for any supermartingale we have that

$$\mathbb{P}(\exists n \in \mathbb{N} : M_n > 1/\delta) \leq \delta \cdot \mathbb{E}[M_0], \tag{13}$$

which is a well-known property of the method of mixtures (de la Peña et al., 2004), we conclude that

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

$\square$

Lemma C.5 (Lemma 19 by Garivier et al., 2018 and comment below)**.** *For all $\lambda, \lambda' \in [0, 1]$ such that either $\lambda \leq \lambda' \leq 1/2$ or $1/2 \leq \lambda' \leq \lambda$, for all real numbers $c \leq 1$,*

$$\log(1 - \lambda c) - \log(1 - \lambda' c) \leq 2|\lambda - \lambda'| \,.$$

C.3. Deviation inequality for bounded distributions

Below, we reproduce the self-normalized Bernstein-type inequality by Domingues et al. (2021c). 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.6 (Bernstein-type concentration inequality)**.** *For all $\delta > 0$,*

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

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

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

C.4. Deviation inequality for Dirichlet distribution

Below we provide the Bernstein-type inequality for weighted sum of Dirichlet distribution, using a result on upper bound on tails of Dirichlet boundary crossing (see Lemma D.1).

Lemma C.7. *For any $p \in \Delta_m$, $f\colon \{0,\dots,m\} \to [0,b]$ such that $f(0) = b$, $p_0 > 0$, and $\mu \in (pf, b)$ there exists a measure $q \in \Delta_m$ such that $p \ll q$, $qf = \mu$ and $\mathcal{K}_{\text{inf}}(p,\mu,f) = \mathrm{KL}(p,q)$.*

Proof. By the variational form of $\mathcal{K}_{\text{inf}}$ (Lemma C.2)

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

Note that $\mathbb{P}(f(X)=b) > 0$ implies $\lambda^\star < 1$. Jensen's inequality and $\mu > pf$ imply $\lambda^\star > 0$. It is easy to check that $\lambda^\star$ satisfies

$$\mathbb{E}\left[\frac{1}{1-\lambda^\star(f(X)-\mu)/(b-\mu)}\right] = \sum_{j=0}^m \frac{p(j)}{1-\lambda^\star(f(j)-\mu)/(b-\mu)} = 1,$$

and

$$\mathbb{E}\left[\frac{f(X)-\mu}{1-\lambda^\star(f(X)-\mu)/(b-\mu)}\right] = \sum_{j=0}^m \frac{p(j)(f(j)-\mu)}{1-\lambda^\star(f(j)-\mu)/(b-\mu)} = 0. \tag{14}$$

Define $q(j) = \frac{p(j)}{1-\lambda^\star(f(j)-\mu)/(b-\mu)}, j = 0,\dots,m$, and let $q = (q_0,\dots,q_m)$. Clearly, $q \in \Delta_m$, $qf = \mu$ by (14) and $p \ll q$. Moreover,

$$\mathcal{K}_{\text{inf}}(p,\mu,f) = \mathbb{E}_{X\sim p}\left[\log\left(1-\lambda^\star\frac{f(X)-\mu}{b-\mu}\right)\right] = \mathbb{E}_p\left[\log\frac{\mathrm{d}p}{\mathrm{d}q}\right] = \mathrm{KL}(p,q).$$

□

Lemma C.8. *For any $\alpha = (\alpha_0,\alpha_1,\dots,\alpha_m) \in \mathbb{N}^{m+1}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0,\dots,m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Then for any $f\colon \{0,\dots,m\} \to [0,b]$ such that $f(0) = b$ and $\delta \in (0,1)$*

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

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

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

Note that such μ exists. Indeed, it follows from the continuity of $\mathcal{K}_{\text{inf}}$ w.r.t. the second argument, see Honda and Takemura (2010, Theorem 7). By 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 Lemma D.1

$$\mathbb{P}_{w \sim \mathcal{D}\text{ir}(\alpha)}[wf \geq qf] = \mathbb{P}_{w \sim \mathcal{D}\text{ir}(\alpha)}[wf \geq \mu] \leq \exp(-\overline{\alpha}\, \mathcal{K}_{\text{inf}}(\overline{p}, \mu, f)) = \delta. \tag{15}$$

By Lemma E.2

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

By Lemma E.3, $\mathrm{Var}_q(f) \leq 2\mathrm{Var}_{\overline{p}}(f) + 4b^2\, \mathrm{KL}(\overline{p}, q)$. The last two inequalities and (15) imply that

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

□

D. Dirichlet Boundary Crossing

In this section we will provide upper and lower bounds on the Dirichlet boundary crossing. The proof of the upper bound follows Baudry et al. (2021); see also Riou and Honda (2020).

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

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

Proof. First if $\mu \leq \overline{p}f$ then the upper bound is trivial since in this case $\mathcal{K}_{\mathrm{inf}}(\overline{p}, \mu, f) = 0$. Assume that $\mu > \overline{p}f$. It is well know fact that $w \sim \mathcal{D}\mathrm{ir}(\alpha)$ may be represented as follows

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

where $Y_\ell \overset{\mathrm{ind}}{\sim} \Gamma(\alpha_\ell, 1), \ell = 0, \dots, m$ and $V_m = \sum_{\ell=0}^m Y_\ell$. Furthermore, denoting $v_\ell \overset{\mathrm{i.i.d}}{\sim} \mathcal{E}(1), \ell = 1, \dots, \overline{\alpha}$, we get

$$wf = \sum_{\ell=0}^m w_\ell f(\ell) = \frac{\sum_{j=1}^{\overline{\alpha}} v_j x_j}{\sum_{j=1}^{\overline{\alpha}} v_j},$$

where $x_j = f(\ell)$ iff $\sum_{k=0}^{\ell} \alpha_k < j \leq \sum_{k=0}^{\ell+1} \alpha_k$. Changing measure and using variational formula for the minimal Kullback-Leibler divergence we get for $\lambda \in [0, 1/(b-\mu))$

$$\begin{aligned}
\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] &= \mathbb{E}_{v_\ell \overset{\mathrm{i.i.d}}{\sim} \mathcal{E}(1)}\left[\mathbb{1}\left\{\sum_{\ell=1}^{\overline{\alpha}} v_\ell(x_\ell - \mu) \geq 0\right\}\right] \\
&= \mathbb{E}_{\hat{v}_\ell \overset{\mathrm{ind}}{\sim} \mathcal{E}\left(1-\lambda(x_\ell-\mu)\right)}\left[\mathbb{1}\left\{\sum_{\ell=1}^{\overline{\alpha}} \hat{v}_\ell(x_\ell - \mu) \geq 0\right\} \cdot \prod_{\ell=1}^{\overline{\alpha}} \frac{\mathrm{e}^{(1-\lambda(x_\ell-\mu))\hat{v}_\ell - \hat{v}_\ell}}{1-\lambda(x_\ell-\mu)}\right] \\
&= \mathrm{e}^{-\sum_{\ell=1}^{\overline{\alpha}} \log(1-\lambda(x_\ell-\mu))} \mathbb{E}_{\hat{v}_\ell \overset{\mathrm{ind}}{\sim} \mathcal{E}\left(1-\lambda(x_\ell-\mu)\right)}\left[\mathbb{1}\left\{\sum_{\ell=1}^{\overline{\alpha}} \hat{v}_\ell(x_\ell - \mu) \geq 0\right\} \mathrm{e}^{-\lambda\sum_{\ell=1}^{\overline{\alpha}} \hat{v}_\ell(x_\ell-\mu))}\right] \\
&\leq \exp\left(-\sum_{\ell=1}^{\overline{\alpha}} \log(1-\lambda(x_\ell-\mu))\right) = \exp\left(-\sum_{\ell=0}^{m} \alpha_\ell \log(1-\lambda(f(\ell)-\mu))\right),
\end{aligned}$$

where the last equality follows from regrouping all x_j back to $f(\ell)$. Since the previous inequality is true for all $\lambda \in [0, 1/(b-\mu))$, then the variational formula (Lemma C.2) allows to conclude

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \leq \exp\left(-\sup_{\lambda\in[0,1/(b-\mu))} \sum_{\ell=0}^{m} \alpha_\ell \log(1-\lambda(f(\ell)-\mu))\right) = \exp(-\overline{\alpha}\,\mathcal{K}_{\mathrm{inf}}(\overline{p}, \mu, f)).$$

□

Theorem D.2 (Lower bound). *For any $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_m) \in \mathbb{N}^{m+1}$ define $\overline{p} \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell/\overline{\alpha}, \ell = 0, \dots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$. Assume that*

$$\alpha_0 \geq \max\left\{\frac{1}{(\sqrt{2\pi}-1)^2} \cdot \left(\frac{2\sqrt{2}}{\sqrt{\log(17/16)}} + \frac{98\sqrt{6}}{9}\right)^2, \frac{\log(10\pi\cdot\overline{\alpha})}{\log(17/16)}\right\}$$

and $\overline{\alpha} \geq 2\alpha_0$. Then for any $f\colon \{0,\dots,m\} \to [0,b_0]$ such that $f(0) = b_0$, $f(j) \leq b < b_0/2, j \in \{1,\dots,m\}$ and $\mu \in (\overline{p}f, b_0)$

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \geq \mu] \geq \exp(-\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p}, \mu, f) - 3/2\log\overline{\alpha}).$$

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

D.1. Proof of Theorem D.2

Throughout this section we will often use the following notations. Let $\mathrm{F}_m(b) = \{f : \{0, \ldots, m\} \to [0, b]\}$ and for $b < b_0$, $\mathrm{F}_m(b_0, b) = \{f : \{0, \ldots, m\} \to [0, b_0], f(0) = b_0, f(j) \le b, j = 1, \ldots, m\}$. For any $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{N}^{m+1}$ define $\overline{p} = \overline{p}(\alpha) \in \Delta_m$ such that $\overline{p}(\ell) = \alpha_\ell / \overline{\alpha}, \ell = 0, \ldots, m$, where $\overline{\alpha} = \sum_{j=0}^m \alpha_j$

Density of weighted sum of the Dirichlet distribution In this section we compute the density of a random variable $Z = wf$, where $w \sim \mathcal{D}\mathrm{ir}(\alpha)$ and $f \in \mathrm{F}_m(b)$.

Proposition D.3. *Let $f \in \mathrm{F}_m(b)$ and $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ such that $\overline{\alpha} = \sum_{j=0}^m \alpha_j > 1$. Assume that Z is not degenerate. Then for any $0 \le u < b_0$*

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

Proof of Proposition D.3 will be given at the end of this paragraph.

A function $g : \mathbb{R}^{m+1} \to \mathbb{R}$ is called a positive homogeneous on a cone $A \subseteq \mathbb{R}^{m+1}$ of degree t if for any $\gamma > 0$ and $x \in A$ we have $g(\gamma x) = \gamma^t g(x)$. Define $\widetilde{\Delta}_m = \mathrm{Conv}(0, \Delta_m) = \{w \in \mathbb{R}_+^{m+1} : \sum_{j=0}^m w_j \le 1\}$ as a pyramid with a base Δ_m and apex at 0. Denote $\overline{\Delta}_m = \{w \in \mathbb{R}^{m+1} : \sum_{\ell=0}^m w_\ell = 1\}$. For $r > 0$ we write $\Delta_m(r) = \{w \in \mathbb{R}_+^{m+1} : \sum_{\ell=0}^m w_\ell = r\}$. Then, clearly $\Delta_m = \Delta_m(1)$. For any $a \in \mathbb{R}^{m+1}$ define $\mathrm{H}_a = \{w \in \mathbb{R}^{m+1} : \langle a, w \rangle = 0\}$.

For any measurable set $A \subseteq \mathbb{R}^{m+1}$ of dimension $n < m + 1$ and any function $g : \mathbb{R}^{m+1} \to \mathbb{R}$ define

$$I_n(g, A) = \int_A g(w) \mathcal{H}^n(\mathrm{d}w),$$

where $\mathcal{H}^n$ is an n-dimensional Hausdorff measure (see Evans and Garzepy, 2018, for definition). If $A = \mathcal{L}(Y)$ for a linear map $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^{m+1}$ and $Y \subseteq \mathbb{R}^n$, then we can write

$$I_n(g, \mathcal{L}(Y)) = [\mathcal{L}] \cdot \int_Y g(\mathcal{L}(y)) \lambda_n(\mathrm{d}y),$$

where λ_n is an n-dimensional Lebesgue measure on Y and $[\mathcal{L}]$ is a Jacobian of the map $\mathcal{L}$ that could be computed as $[\mathcal{L}] = \sqrt{\det(\mathcal{L}\mathcal{L}^\top)}$. Let us define an affine map $\mathcal{L}_a^t : \mathbb{R}^m \to \mathbb{R}^{m+1}$ that transforms $\mathbb{R}^m$ to H_a^t by mapping $x_1, \ldots, x_m$ to $w_1, \ldots, w_m$ and $w_0 = \frac{t - \sum_{j=1}^m a_j x_j}{a_0}$ for $a_0 > 0$ (without loss of generality). The linear part of this map has a Jacobian that is equal to $[\mathcal{L}_a^t] = \frac{\|a\|_2}{a_0}$ (see Lemma E.6). Additionally, define $\mathcal{L}_a = \mathcal{L}_a^0$.

Lemma D.4. *Let g be a positively homogeneous function of degree $t > -m$ on $\mathbb{R}_{++}^{m+1}$. Then we have*

$$I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a) = \frac{\mathrm{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0)}{m + t} I_{m-1}(g, \Delta_m \cap \mathrm{H}_a).$$

Proof. Denote $h = \mathrm{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0)$. The change of variable formula (Evans and Garzepy, 2018, 3.4.3) implies that

$$I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a) = \int_0^h I_{m-1}(g, \Delta_m(s/h) \cap \mathrm{H}_a) \mathrm{d}s.$$

Using definition of a positive homogeneous function and properties of the Haudorff measure $\mathcal{H}^{m-1}$, we derive

$$\begin{aligned} I_{m-1}(g, \Delta_m(s/h) \cap \mathrm{H}_a) &= \int_{\Delta_m \cap \mathrm{H}_a} g(w \cdot (s/h)) \mathcal{H}^{m-1}(\mathrm{d}(w \cdot s/h)) = \\ &= \left(\frac{s}{h}\right)^{m+t-1} \int_{\Delta_m \cap \mathrm{H}_a} g(w) \mathcal{H}^{m-1}(\mathrm{d}w) = \left(\frac{s}{h}\right)^{m+t-1} \cdot I_{m-1}(g, \Delta_m \cap \mathrm{H}_a). \end{aligned}$$

Hence,

$$I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a) = \frac{I_{m-1}(g, \Delta_m \cap \mathrm{H}_a)}{h^{m+t-1}} \int_0^h s^{m+t-1} \mathrm{d}s = \frac{h I_{m-1}(g, \Delta_m \cap \mathrm{H}_a)}{m + t}.$$

□

Now we see that in order to find $I_{m-1}(g, \Delta_m \cap \mathrm{H}_a)$ it is sufficient to compute the integral $I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a)$ and the distance $\mathrm{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0)$. This distance was computed in Dirksen (2015).

Lemma D.5 (Lemma 3.2 in Dirksen, 2015)**.** *Let $a \in \mathbb{R}^{m+1}$ be such that $\|a\|_2 = 1$. Then*

$$\mathrm{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0) = \frac{1}{\sqrt{m+1-(\sum_{i=0}^m a_i)^2}}.$$

Without normalization of the vector a we get as a corollary the following representation.

Corollary D.6. *Let $a \in \mathbb{R}^{m+1}$. Then*

$$\mathrm{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0) = \frac{\|a\|_2}{\sqrt{(m+1)\left(\sum_{j=0}^m a_j^2\right) - (\sum_{j=0}^m a_j)^2}}.$$

Next we provide another representation of the integral $I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a)$. We follow Lasserre (2020) and use the same technique based on the Laplace transform.

Lemma D.7. *Let g be a positively homogeneous function of degree t on $\mathbb{R}^{m+1}_{++}$ such that $t > -(1+m)$ and $\int_{\widetilde{\Delta}_m} |g(w)| \mathrm{d}w < \infty$. Then*

$$I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a) = \frac{1}{\Gamma(1+m+t)} \int_{\mathbb{R}^{m+1}_+ \cap \mathrm{H}_a} g(w) \exp\left(-\sum_{\ell=0}^m w_\ell\right) \mathcal{H}^m(\mathrm{d}w).$$

Proof. Consider $h(y) = \int_{w \geq 0, \langle \mathbf{1}, w \rangle \leq y, \langle a, w \rangle = 0} g(w) \mathcal{H}^m(\mathrm{d}w)$. Clearly, $h(1) = I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a)$. Since g is positively homogeneous function we get $h(y) = y^{m+t} h(1)$. This implies that the Laplace transform of h is equal to $L(\lambda) = \int_0^\infty h(y) \mathrm{e}^{-\lambda y} \mathrm{d}y = h(1) \frac{\Gamma(m+t+1)}{\lambda^{m+t+1}}$. On the other hide, the Laplace transform $L(\lambda)$ may be calculated via a linear parametrization of the subspace $\langle a, w \rangle = 0$ using the map $\mathcal{L}_a$ and the Fubini theorem

$$\begin{aligned}
L(\lambda) &= \int_0^\infty \mathrm{e}^{-\lambda y} \left[\int_{w \in \mathbb{R}^{m+1}_+, \langle \mathbf{1}, w \rangle \leq y, \langle a, w \rangle = 0} g(w) \mathcal{H}^m(\mathrm{d}w)\right] \mathrm{d}y \\
&= [\mathcal{L}_a] \int_{\mathcal{L}_a(x) \geq 0} \mathrm{d}x \cdot g(\mathcal{L}_a(x)) \cdot \left[\int_{\langle \mathbf{1}, \mathcal{L}_a(x) \rangle \leq y} \mathrm{e}^{-\lambda y} \mathrm{d}y\right] \\
&= \frac{[\mathcal{L}_a]}{\lambda} \int_{\mathcal{L}_a(x) \geq 0} g(\mathcal{L}_a(x)) \exp(-\lambda \langle \mathbf{1}, \mathcal{L}_a(x) \rangle) \mathrm{d}x \\
&= \frac{[\mathcal{L}_a]}{\lambda^{m+t+1}} \int_{\mathcal{L}_a(x) \geq 0} g(\mathcal{L}_a(x)) \exp(-\langle \mathbf{1}, \mathcal{L}_a(x) \rangle) \mathrm{d}x \\
&= \frac{1}{\lambda^{m+t+1}} \int_{\mathbb{R}^{m+1}_+ \cap \mathrm{H}_a} g(w) \exp\left(-\sum_{\ell=0}^m w_\ell\right) \mathcal{H}^m(\mathrm{d}w).
\end{aligned}$$

Identifying two ways of computation of the Laplace transform, we finish the proof. □

We now compute the integral in the r.h.s. of Lemma D.7. We shall use the Fourier transform method and follow the approach of Dirksen (2015).

Lemma D.8. *Let $g(w) = w_0^{\alpha_0 - 1} \cdot \ldots \cdot w_m^{\alpha_m - 1}$. Then we have*

$$\int_{\mathbb{R}^{m+1}_+ \cap \mathrm{H}_a} g(w) \exp\left(-\sum_{i=0}^m w_i\right) \mathcal{H}^m(\mathrm{d}w) = \frac{\|a\|_2 \cdot \prod_{j=0}^m \Gamma(\alpha_j)}{2\pi} \int_{\mathbb{R}} \prod_{j=0}^m (1 + \mathrm{i} a_j \tau)^{-\alpha_j} \mathrm{d}\tau.$$

Proof. Denote for any $t \in \mathbb{R}$

$$G(t) = \int_{\forall j: \langle a, w \rangle = t, w_j \geq 0} g(w) \exp\left(-\sum_{\ell=0}^m w_\ell\right) \mathcal{H}^m(\mathrm{d}w).$$

Next, we apply affine parametrization induced by map $\mathcal{L}_a^t$

$$G(t) = [\mathcal{L}_a^t] \int_{\substack{\forall j\ x_j \geq 0 \\ t - \sum_{j=1}^m a_j x_j \geq 0}} \prod_{j=1}^m x_j^{\alpha_j - 1} \cdot \left(\frac{t - \sum_{j=1}^m a_j x_j}{a_0} \right)^{\alpha_0 - 1} \exp\left(- \sum_{j=1}^m x_j - \left(\frac{t - \sum_{j=1}^m a_j x_j}{a_0} \right) \right) \mathrm{d}x.$$

By Lemma E.6 $[\mathcal{L}_a^t] = \frac{\|a\|_2}{a_0}$. The Fourier transform of G may be calculated using the Fubini's theorem

$$\mathcal{F}[G](\tau) = \frac{\|a\|_2}{\sqrt{2\pi} \cdot a_0} \iint\limits_{\substack{\forall j\ x_j \geq 0 \\ t - \sum_{j=1}^m a_j x_j \geq 0 \\ t \in \mathbb{R}}} \prod_{j=1}^m x_j^{\alpha_j - 1} \left(\frac{t - \sum_{j=1}^m a_j x_j}{a_0} \right)^{\alpha_0 - 1} \exp\left(- \sum_{j=1}^m x_j - \left(\frac{t - \sum_{j=1}^m a_j x_j}{a_0} \right) \right) \cdot \mathrm{e}^{-\mathrm{i}t\tau} \mathrm{d}t \mathrm{d}x.$$

By the change of variables $s = t - \sum_{j=1}^m a_j x_j$, the above integral can be represented as a product of $m+1$ one-dimensional integrals

$$\mathcal{F}[G](\tau) = \frac{\|a\|_2}{\sqrt{2\pi} \cdot a_0} \int_0^\infty \mathrm{d}x_1 x_1^{\alpha_1 - 1} \mathrm{e}^{-x_1 - \mathrm{i}\tau a_1 x_1} \ldots \int_0^\infty \mathrm{d}x_m x_m^{\alpha_m - 1} \mathrm{e}^{-x_m - \mathrm{i}\tau a_m x_m} \int_0^\infty \mathrm{d}s \left(\frac{s}{a_0} \right)^{\alpha_0 - 1} \mathrm{e}^{-s/a_0 - \mathrm{i}\tau s}.$$

Performing the change of variables $x_0 = s/a_0$, we arrive at the product of m characteristic functions of independent $\Gamma(\alpha_\ell, 1)$-distributed random variables

$$\mathcal{F}[G](\tau) = \frac{\|a\|_2}{\sqrt{2\pi}} \cdot \prod_{j=0}^m \Gamma(\alpha_j) \cdot \frac{1}{(1 + \mathrm{i}a_j \tau)^{\alpha_j}}.$$

Finally, by the inverse Fourier transform

$$G(0) = \frac{\|a\|_2}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathcal{F}[G](\tau) \mathrm{d}\tau = \frac{\|a\|_2 \cdot \prod_{j=0}^m \Gamma(\alpha_j)}{2\pi} \int_{\mathbb{R}} \prod_{j=0}^m (1 + \mathrm{i}a_j \tau)^{-\alpha_j} \mathrm{d}\tau.$$

□

Remark D.9. Notice that the value of the integral is the right-hand side is real because the function under integral has even real part and odd imaginary one.

Corollary D.10. *Let $g(w) = \Gamma(\overline{\alpha}) \prod_{j=0}^m w_j^{\alpha_j - 1} / \Gamma(\alpha_j)$, where $\overline{\alpha} > 1$ and $a \in \mathbb{R}^{m+1}$. Then*

$$I_{m-1}(g, \Delta_m \cap \mathrm{H}_a) = (\overline{\alpha} - 1) \sqrt{(m+1) \left(\sum_{i=0}^m a_i^2 \right) - \left(\sum_{i=0}^m a_i \right)^2} \cdot \frac{1}{2\pi} \int_{\mathbb{R}} \prod_{j=0}^m (1 + \mathrm{i}a_j s)^{-\alpha_j} \mathrm{d}s.$$

Proof. Notice that g is positively homogeneous function of degree $\overline{\alpha} - (m+1) > -m$ on $\mathbb{R}_{++}^{m+1}$. Hence, we may apply Lemma D.4 and Lemma D.7. We obtain

$$\begin{aligned} I_{m-1}(g, \Delta_m \cap \mathrm{H}_a) &= \frac{\overline{\alpha} - 1}{\operatorname{dist}(\overline{\Delta} \cap \mathrm{H}_a, 0)} \cdot I_m(g, \widetilde{\Delta}_m \cap \mathrm{H}_a) \\ &= \frac{\overline{\alpha} - 1}{\operatorname{dist}(\overline{\Delta} \cap \mathrm{H}_a, 0) \Gamma(\overline{\alpha})} \int_{\mathbb{R}_+^{m+1} \cap \mathrm{H}_a} g(w) \exp\left(- \sum_{\ell=0}^m w_\ell \right) \mathcal{H}^m(\mathrm{d}w). \end{aligned}$$

The last integral could be computed by Lemma D.8. We have

$$I_{m-1}(g, \Delta_m \cap \mathrm{H}_a) = \frac{\overline{\alpha} - 1}{\operatorname{dist}(\overline{\Delta}_m \cap \mathrm{H}_a, 0)} \cdot \frac{\|a\|_2}{2\pi} \int_{\mathbb{R}} \prod_{j=0}^m (1 + \mathrm{i}a_j s)^{-\alpha_j} \mathrm{d}s.$$

Finally, we apply Corollary D.6 to conclude the statement. □

Now we are ready to prove Proposition D.3.

Proof of Proposition D.3. First, we give a formula for p_Z in terms of I_{m-1}. We start from rewriting the probability in terms of a usual integral

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \le \mu] = \int_{w\ge 0, \sum_{i=1}^m w_i \le 1, wf\le\mu} g\left(1-\sum_{i=1}^m w_i, w_1, \dots, w_m\right)\mathrm{d}w_1, \dots, \mathrm{d}w_m$$

where $g(w) = \Gamma(\overline{\alpha})\prod_{j=0}^m w_j^{\alpha_j-1}/\Gamma(\alpha_j)$ is the density of the Dirichlet distribution. We note that this transform exactly defines a map $\mathcal{L}_{\mathbf{1}}^1$. Then we apply changing of variables formula (Evans and Garzepy, 2018, 3.4.3) using map $\phi(x) = f^\intercal\mathcal{L}_{\mathbf{1}}^1(x)$

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \le \mu] = \frac{1}{[\phi]}\int_0^\mu \left[\int_{\mathcal{L}_{\mathbf{1}}^1(x)\ge 0, f^\intercal\mathcal{L}_{\mathbf{1}}^1(w)=u} g(\mathcal{L}_{\mathbf{1}}^1(x))\mathcal{H}^m(\mathrm{d}x)\right]\mathrm{d}u$$

Define a vector $c = \mathcal{L}_{\mathbf{1}}^\intercal f$. Then we apply changing of variables formula (Evans and Garzepy, 2018, 3.3.3) to the inner integral using parametrization through map $\mathcal{L}_c^u$

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \le \mu] = \frac{1}{[\phi]}\int_0^\mu \left[[\mathcal{L}_c^u]\int_{\mathcal{L}_{\mathbf{1}}^1(\mathcal{L}_c^u w)\ge 0} g(\mathcal{L}_{\mathbf{1}}^1(\mathcal{L}_c^u(z)))\mathrm{d}z\right]\mathrm{d}u$$

We note that a Jacobian of $\mathcal{L}_c^u$ does not depend on the shift parameter u, therefore could me moved from the integral sign. Next we apply changing of variables formula for a map $\mathcal{L}_{\mathbf{1}}^1 \circ \mathcal{L}_c^u$

$$\mathbb{P}_{w\sim\mathcal{D}\mathrm{ir}(\alpha)}[wf \le \mu] = \frac{[\mathcal{L}_c^u]}{[\phi][\mathcal{L}_{\mathbf{1}}^1\circ\mathcal{L}_c^u]}\int_0^\mu \left[\int_{\Delta_m, wf=u} g(w)\mathcal{H}^{m-1}(\mathrm{d}w)\right]\mathrm{d}u.$$

To compute all Jacobians we shall use Lemma E.6 and Lemma E.7, and notice that $[\phi] = \|c\|_2$. As a result, p_Z can be represented as the following integral

$$\begin{aligned}
p_Z(u) &= \frac{1}{\sqrt{(m+1)\sum_{j=0}^m f^2(j) - \left(\sum_{j=0}^m f(j)\right)^2}}\int_{\Delta_m\cap \mathrm{H}_f^u} g(w)\mathcal{H}^{m-1}(\mathrm{d}w) \\
&= \frac{1}{\sqrt{(m+1)\sum_{j=0}^m f^2(j) - \left(\sum_{j=0}^m f(j)\right)^2}} I_{m-1}(g, \Delta_m\cap\mathrm{H}_f^u),
\end{aligned}$$

where $\mathrm{H}_f^u = \{w\in\mathbb{R}^{m+1} : wf = u\}$. Unfortunately, we cannot apply the previous result directly because the hyperplane H_f^u does not intersect 0 in general. To overcome this issue, define the following vector $a(u)_j = f(j) - u$. Note that $\langle w, a(u)\rangle = 0$ iff $\langle w, f - u\mathbf{1}\rangle = wf - u = 0$, where we used $\langle w, \mathbf{1}\rangle = 1$. Hence $\mathrm{H}_f^u\cap\Delta_m = \mathrm{H}_{a(u)}\cap\Delta_m$. We can apply Corollary D.10 to the subspace $\mathrm{H}_{a(u)}$

$$\begin{aligned}
I_{m-1}(g, \Delta_m\cap\mathrm{H}_f^u) &= \frac{\overline{\alpha}-1}{2\pi}\sqrt{(m+1)\left(\sum_{j=0}^m a(u)_j^2\right) - \left(\sum_{j=0}^m a(u)_j\right)^2}\cdot\int_{\mathbb{R}}\prod_{j=0}^m (1+\mathrm{i}a(u)_j s)^{-\alpha_j}\mathrm{d}s \\
&= \frac{\overline{\alpha}-1}{2\pi}\sqrt{(m+1)\sum_{j=0}^m (f(j)-u)^2 - \left(\sum_{j=0}^m (f(j)-u)\right)^2}\cdot\int_{\mathbb{R}}\prod_{j=0}^m (1+\mathrm{i}(f(j)-u)t)^{-\alpha_j}\mathrm{d}t.
\end{aligned}$$

Finally, we will rewrite the expression under square root as follows

$$
\begin{aligned}
(m+1)\sum_{j=0}^{m}(f(j)-u)^2 - \left(\sum_{j=0}^{m}(f(j)-u)\right)^2 &= (m+1)\left(\sum_{j=0}^{m}f(j)^2 - 2u\sum_{j=0}^{m}f(j) + (m+1)u^2\right) \\
&\quad - \left(\sum_{j=0}^{m}f(j)\right)^2 + 2u(m+1)\sum_{j=0}^{m}f(j) - (m+1)^2u^2 \\
&= (m+1)\sum_{j=0}^{m}f(j)^2 - \left(\sum_{j=0}^{m}f(j)\right)^2.
\end{aligned}
$$

We conclude the proof of proposition. ◻

Saddle point method In this paragraph, we analyze the asymptotic behavior of the density of a linear statistic of the Dirichlet distribution using the method of saddle point (Olver, 1997; Fedoryuk, 1977) and obtain sharp bounds on remainder terms.

Proposition D.11. *Let $f \in \mathrm{F}_m(b_0, b)$ and let $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ be a fixed vector with $\alpha_0 \geq 2$. Then for any $u \in (\overline{p}f, b_0)$,*

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

where

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

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

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

Remark D.12. From these bounds on remainder terms we see that α_0 should increase at least as $\log\overline{\alpha}$ in order to make R_3 small enough.

Proof. Let us first rewrite our integral in the form

$$
\begin{aligned}
I = \int_{\mathbb{R}} \prod_{j=0}^{n}(1+\mathrm{i}(f(j)-u)s)^{-\alpha_j}\,\mathrm{d}s &= \int_{\mathbb{R}} \exp\left(-\overline{\alpha}\sum_{j=0}^{m}\overline{p}_j\log(1+\mathrm{i}(f(j)-u)s)\right)\mathrm{d}s \\
&= \int_{\mathbb{R}} \exp(-\overline{\alpha}\,\mathbb{E}_{X\sim\overline{p}}[\log(1+\mathrm{i}(f(X)-u)s)])\,\mathrm{d}s, \qquad (16)
\end{aligned}
$$

where we choose the principle branch of the complex logarithmic function. Denote $S(z) = \mathbb{E}_{X\sim\overline{p}}[\log(1+\mathrm{i}(f(X)-u)z)]$. In the sequel we shall write for simplicity $\mathbb{E}$ instead of $\mathbb{E}_{X\sim\overline{p}}$.

Since $f(X) \le b_0$, this function is holomorphic for $|\operatorname{Im} z| < 1/(b_0 - u)$ and $\operatorname{Re} z \in \mathbb{R}$. The last integral representation (16) allows to use the method of saddle point (Olver, 1997). Next, we are going to compute the saddle points of the function S. To do it, compute the derivative of the function S at complex point $z = x + \mathrm{i}y$

$$S'(z) = \mathrm{i}\mathbb{E}\left[\frac{f(X)-u}{1+\mathrm{i}(f(X)-u)z}\right] = \mathbb{E}\left[\frac{x(f(X)-u)^2 + \mathrm{i}(f(X)-u)(1-y(f(X)-u))}{(1-y(f(X)-u))^2 + x^2(f(X)-u)^2}\right] = 0.$$

Notice that the real part of the expression above is zero if and only if $x = 0$. Therefore the saddle points could be only on the imaginary line $\mathrm{i}\mathbb{R}$. They can be found from the equation

$$S'(\mathrm{i}y) = \mathrm{i}\mathbb{E}\left[\frac{f(X)-u}{1-y(f(X)-u)}\right] = 0.$$

Note that for $y \ge 0$, this equation coincides with the optimality condition for $\lambda^\star$ in the definition of $\mathcal{K}_{\text{inf}}(\overline{p}, u, f)$. Since $\overline{p}f < u < b_0$, the function $y \mapsto S(\mathrm{i}y) = \mathbb{E}[\log(1 - y(f(X)-u))]$ is strictly concave in y and, therefore, equation $S'(\mathrm{i}y) = 0$ has a unique solution $y = \lambda^\star$. Thus the unique saddle point of S is equal to $z_0 = \mathrm{i}\lambda^\star$. Next, let us change the integration contour to $\gamma^\star = \mathbb{R} + \mathrm{i}\lambda^\star$. To prove that this contour is suitable, let us show that the real part of S achieves a minimum at z_0 over all $z \in \gamma^\star$

$$\operatorname{Re} S(x + \mathrm{i}\lambda^\star) = \frac{1}{2}\mathbb{E}\big[\log\big((1-\lambda^\star(f(X)-u))^2 + x^2(f(X)-u)^2\big)\big].$$

The minimum of $\operatorname{Re} S(x + \mathrm{i}\lambda^\star)$ is achieved for $x = 0$, therefore the contour $\gamma^\star$ is suitable. Hence, we can apply the Laplace method after a simple change of coordinates

$$I = \int_{\mathbb{R}} \exp(-\overline{\alpha}\,\mathbb{E}[\log(1-\lambda^\star(f(X)-u) + \mathrm{i}s(f(X)-u))])\,\mathrm{d}s$$

Denote

$$T(s) = \mathbb{E}[\log(1-\lambda^\star(f(X)-u) + \mathrm{i}s(f(X)-u))].$$

Fix a cut-off parameter $K > 0$ and define $\kappa_1 = T(-K) - T(0)$, $\kappa_2 = T(K) - T(0)$. Next, similarly to 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 $q_1(v_1) = \frac{1}{T'(-s)}$, $q_2(v_2) = \frac{1}{T'(s)}$. Using the first order Taylor expansion, we can write $q_1(v_1) = \frac{1}{\sqrt{2T''(0)\cdot v_1}} + r_1(v_1)$, $q_2(v_2) = \frac{1}{\sqrt{2T''(0)\cdot v_2}} + r_2(v_2)$. Then we have the following decomposition

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

where

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

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. Our next goal is to analyze these remainder terms and compute upper bounds on their absolute values.

Term R_2. First, we derive the uniform bounds for r_2 and r_1. Using the expansions

$$\begin{aligned}
T(s) &= T(0) + T'(0)\cdot s + \frac{T''(0)}{2}s^2 + \frac{T'''(\xi_1)}{6}s^3, & \xi_1 \in (0, s)\\
T'(s) &= T'(0) + T''(0)s + \frac{T'''(\xi_2)}{2}s^2, & \xi_2 \in (0, s).
\end{aligned}$$

and noting that $T'(0)=0$, we get

$$
\begin{aligned}
|r_2(v)| &= \left|\frac{1}{T'(s)} - \frac{1}{\sqrt{2T''(0)(T(s)-T(0))}}\right| \\
&= \left|\frac{\sqrt{T''(0)^2 s^2 + T''(0)\frac{T'''(\xi_1)}{3}s^3} - T''(0)s - \frac{T'''(\xi_2)}{2}s^2}{[T''(0)s + \frac{T'''(\xi_2)}{2}s^2]\cdot\sqrt{T''(0)^2 s^2 + T''(0)\frac{T'''(\xi_1)}{2}s^3}}\right| \\
&= \left|T''(0)\frac{\sqrt{1+\frac{T'''(\xi_1)}{3T''(0)}s} - 1 - \frac{T'''(\xi_2)}{2T''(0)}s}{s\cdot[T''(0) + \frac{T'''(\xi_2)}{2}s]\cdot\sqrt{T''(0)^2 + T''(0)\frac{T'''(\xi_1)}{3}s}}\right|.
\end{aligned}
$$

Next by applying the inequality $\sqrt{1+x}-1 = \frac{x}{2} - \frac{x^2}{8(1+\xi_3)^{3/2}}$ for $|\xi_3| < x$,

$$
\begin{aligned}
|r_2(v)| &= \left|T''(0)\frac{\frac{T'''(\xi_1)}{6T''(0)} - \frac{T'''(\xi_2)}{2T''(0)} - s\cdot\frac{(T'''(\xi_1)/T''(0))^2}{8\cdot 9\cdot(1+\xi_3)^{3/2}}}{[T''(0) + \frac{T'''(\xi_2)}{2}s]\cdot\sqrt{T''(0)^2 + T''(0)\frac{T'''(\xi_1)}{3}s}}\right| \\
&= \left|\frac{\frac{T'''(\xi_1)}{6} - \frac{T'''(\xi_2)}{2} - s\cdot\frac{(T'''(\xi_1))^2}{8\cdot 9\cdot(1+\xi_3)^{3/2}\cdot T''(0)}}{[T''(0) + \frac{T'''(\xi_2)}{2}s]\cdot\sqrt{T''(0)^2 + T''(0)\frac{T'''(\xi_1)}{3}s}}\right| \le \frac{\frac{2}{3}\sup_{u\in(0,s)}|T'''(u)| + s\cdot\frac{\sup_{u\in(0,s)}|T'''(u)|^2}{72\cdot|T''(0)|}}{|T''(0) + \frac{T'''(\xi_2)}{2}s|\cdot\sqrt{|T''(0)^2 + T''(0)\frac{T'''(\xi_1)}{3}s|}}.
\end{aligned}
$$

Let us analyze the second and third derivative of T

$$
T''(s) = \mathbb{E}\left[\left(\frac{f(X)-u}{1-\lambda^\star(f(X)-u) + is(f(X)-u)}\right)^2\right], \quad T'''(s) = -2i\mathbb{E}\left[\left(\frac{f(X)-u}{1-\lambda^\star(f(X)-u) + is(f(X)-u)}\right)^3\right].
$$

Define a random variable $Y_s = \frac{f(X)-u}{1-\lambda^\star(f(X)-u)+is(f(X)-u)}$, then $T''(s) = \mathbb{E}[Y_s^2]$, $T'''(s) = -2i\mathbb{E}[Y_s^3]$. Let us compute an upper bound on the absolute value of $T'''(s)$

$$
|T'''(s)| \le 2\mathbb{E}\left[\frac{|f(X)-u|^3}{((1-\lambda^\star(f(X)-u))^2 + s^2(f(X)-u)^2)^{3/2}}\right] \le 2\mathbb{E}[|Y_0|^3].
$$

By choosing $1/(2K) = \max\left\{\frac{b_0-u}{1-\lambda^\star(b_0-u)}, \frac{u}{1+\lambda^\star u}\right\}$, we ensure that $\mathbb{E}[Y_0^2] - s\mathbb{E}[|Y_0|^3] \ge \frac{1}{2}\mathbb{E}[Y_0^2]$ for all $0 \le s < K$, since

$$
\mathbb{E}[|Y_0|^3] \le \max_{j\in\{0,\ldots,m\}} \frac{|f(j)-u|}{1-\lambda^\star(f(j)-u)}\mathbb{E}[Y_0^2] \le \max\left\{\frac{b_0-u}{1-\lambda^\star(b_0-u)}, \frac{u}{1+\lambda^\star u}\right\}\mathbb{E}[Y_0^2] \le \frac{1}{2K}\mathbb{E}[Y_0^2].
$$

Hence

$$
\begin{aligned}
|r_2(v)| &\le \frac{\frac{4}{3}\mathbb{E}[|Y_0|^3] + s\frac{\mathbb{E}[|Y_0|^3]^2}{18\cdot\mathbb{E}[Y_0^2]}}{(\mathbb{E}[Y_0^2] - \mathbb{E}[|Y_0|^3]s)\cdot\sqrt{\mathbb{E}[Y_0^2]}\cdot\sqrt{\mathbb{E}[Y_0^2] - \mathbb{E}[|Y_0|^3]\frac{2s}{3}}} \\
&\le \frac{4/3 + 1/36}{1/2\cdot\sqrt{2/3}}\cdot\frac{\mathbb{E}[|Y_0|^3]}{\mathbb{E}[Y_0^2]^2} \le \frac{49\sqrt{6}}{36\mathbb{E}[Y_0^2]}\cdot\max\left\{\frac{b_0-u}{1-\lambda^\star(b_0-u)}, \frac{u}{1+\lambda^\star u}\right\}.
\end{aligned}
$$

Next, using the bound

$$
\mathbb{E}[Y_0^2] = \frac{\alpha_0}{\overline{\alpha}}\left(\frac{b_0-u}{1-\lambda^\star(b_0-u)}\right)^2 + \sum_{j=1}^m \frac{\alpha_j}{\overline{\alpha}}\left(\frac{f(j)-u}{1-\lambda^\star(f(j)-u)}\right)^2 \ge \frac{\alpha_0}{\overline{\alpha}}\left(\frac{b_0-u}{1-\lambda^\star(b_0-u)}\right)^2,
$$

we obtain

$$
|r_2(v)| \le \frac{49\sqrt{6}}{36\sqrt{\mathbb{E}[Y_0^2]}}\cdot\sqrt{\frac{\overline{\alpha}}{\alpha_0}}\max\left\{1, \frac{u(1-\lambda^\star(b_0-u))}{(1+\lambda^\star u)(b_0-u)}\right\}.
$$

Next we use Lemma 12 from (Honda and Takemura, 2010)

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

thus

$$|r_2(v)| \leq \frac{49\sqrt{6}}{36\sqrt{\mathbb{E}[Y_0^2]}} \cdot \sqrt{\frac{\overline{\alpha}}{\alpha_0}} \frac{b_0}{b_0 - \overline{p}f}.$$

A similar bound also holds for $r_1(v)$ by symmetry. Set $c_2' = \frac{49\sqrt{6}}{18}$, then

$$|R_2(\alpha)| \leq \frac{c_2'}{2\sqrt{\mathbb{E}[Y_0^2]}} \cdot \sqrt{\frac{\overline{\alpha}}{\alpha_0}} \cdot \left| \int_0^{\kappa_2} e^{-\overline{\alpha} v} \mathrm{d}v + \int_0^{\kappa_1} e^{-\overline{\alpha} v} \mathrm{d}v \right| \leq \frac{c_2'}{\sqrt{\mathbb{E}[Y_0^2]}} \cdot \frac{1 + \exp(-\overline{\alpha}\kappa)}{\sqrt{\overline{\alpha} \cdot \alpha_0}} \frac{b_0}{b_0 - \overline{p}f},$$

where $\kappa = \min\{\operatorname{Re}\kappa_1, \operatorname{Re}\kappa_2\}$. Using the identity

$$\operatorname{Re}\kappa_1 = \operatorname{Re}\kappa_2 = \frac{1}{2}\mathbb{E}\left[\log\left(\frac{(1 - \lambda^\star(f(X) - u))^2 + K^2(f(X) - u)^2}{(1 - \lambda^\star(f(X) - u))^2}\right)\right] = \frac{1}{2}\mathbb{E}\left[\log\left(1 + K^2 Y_0^2\right)\right]$$

and the inequality

$$\begin{aligned}
\mathbb{E}\left[\log\left(1 + K^2 Y_0^2\right)\right] &= \frac{\alpha_0}{\overline{\alpha}} \log\left(1 + K^2 \cdot \left(\frac{b_0 - u}{1 - \lambda^\star(b_0 - u)}\right)^2\right) + \sum_{j=1}^m \frac{\alpha_j}{\overline{\alpha}} \log\left(1 + K^2 \cdot \left(\frac{f(j) - u}{1 - \lambda^\star(f(j) - u)}\right)^2\right) \\
&\geq \frac{\alpha_0}{\overline{\alpha}} \log\left(1 + K^2 \cdot \left(\frac{b_0 - u}{1 - \lambda^\star(b_0 - u)}\right)^2\right) \geq \frac{\alpha_0}{\overline{\alpha}} \log\left(1 + \frac{1}{4}\left(\frac{b_0 - \overline{p}f}{b_0}\right)^2\right),
\end{aligned}$$

we have $\kappa = \operatorname{Re}\kappa_2 = \operatorname{Re}\kappa_1 \geq c_\kappa \cdot \frac{\alpha_0}{\overline{\alpha}}$ with $c_\kappa = 1/2 \cdot \log\left(1 + \frac{1}{4}\left(\frac{b_0 - \overline{p}f}{b_0}\right)^2\right)$. Since $\alpha_0 \geq 2$, we also have $\exp(-c_\kappa \alpha_0) \leq 1$. Finally setting $c_2 = 2 \cdot c_2' = \frac{49\sqrt{6}}{9}$, we derive the following bound on R_2

$$|R_2(\alpha)| \leq \frac{c_2}{\sqrt{\mathbb{E}[Y_0^2]}} \cdot \frac{1}{\sqrt{\overline{\alpha}}\,\alpha_0} \cdot \frac{b_0}{b_0 - \overline{p}f}.$$

Term R_1. By Theorem 2.4 of Borwein and Chan (2007) we have the following bound on complex gamma function for any complex z with $\operatorname{Re} z > 0$

$$\left|\Gamma\left(\frac{1}{2}, z\right)\right| \leq \frac{2e^{-\operatorname{Re} z}}{|z|^{1/2}}.$$

Therefore,

$$|R_1(\alpha)| \leq \frac{4}{\sqrt{2T''(0)|\kappa|}} \cdot \frac{\exp(-\overline{\alpha}\,\kappa)}{\overline{\alpha}}.$$

Set $c_1 = 2\sqrt{2}$, then we have under our choice of K and κ,

$$|R_1(\alpha)| \leq \frac{c_1}{\sqrt{\mathbb{E}[Y_0^2] c_\kappa \alpha_0}} \cdot \frac{\exp(-c_\kappa \alpha_0)}{\overline{\alpha}^{1/2}}.$$

Term R_3. We have

$$\left|\int_K^\infty \exp(-\overline{\alpha} T(s))\,\mathrm{d}s\right| \leq \exp(-\overline{\alpha} \cdot \operatorname{Re}[T(K) - T(0)]) \cdot \exp(-\overline{\alpha} T(0)) \cdot \int_K^\infty \exp(-\overline{\alpha} \operatorname{Re}[T(s) - T(K)])\,\mathrm{d}s. \quad (17)$$

Our goal is to bound the last integral. Let us analyze the function under exponent after a change of variables $s \to t + K$

$$\begin{aligned}
q(t) \triangleq \operatorname{Re}[T(t+K) - T(K)] &= \frac{1}{2}\mathbb{E}\left[\log\left(\frac{(1 - \lambda^\star(f(X) - u))^2 + (t+K)^2(f(X) - u)^2}{(1 - \lambda^\star(f(X) - u))^2 + K^2(f(X) - u)^2}\right)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\log\left(1 + \frac{(2tK + t^2)(f(X) - u)^2}{(1 - \lambda^\star(f(X) - u))^2 + K^2(f(X) - u)^2}\right)\right].
\end{aligned}$$

Define a function $g(j) = \frac{(f(j)-u)^2}{(1-\lambda^\star(f(j)-u))^2 + K^2(f(j)-u)^2} \geq 0$. Then

$$q(t) = \frac{1}{2}\mathbb{E}\big[\log\big(1 + (2tK + t^2)g(X)\big)\big] \geq \frac{1}{2} \cdot \frac{\alpha_0}{\overline{\alpha}} \cdot \log(1 + (2tK + t^2)g(0)), \tag{18}$$

by positivity of $g(j)$. By choosing $b_0 > u$, we have $g(0) > 0$, therefore (18) is a non-trivial lower bound. By substitution of (18) into the integral (17), we get

$$\begin{aligned}\int_K^\infty \exp(-\overline{\alpha}\operatorname{Re}[T(s) - T(K)])\mathrm{d}s &= \int_0^\infty \exp(-\overline{\alpha}q(t))\mathrm{d}t \leq \int_0^\infty \left(\frac{1}{1 + (2tK + t^2)g(0)}\right)^{\alpha_0/2} \mathrm{d}t \\ &\leq \int_0^\infty \left(\frac{1}{1 + t^2 g(0)}\right)^{\alpha_0/2} \mathrm{d}t = \frac{\sqrt{\pi}\cdot\Gamma\left(\frac{\alpha_0-1}{2}\right)}{2\sqrt{g(0)}\cdot\Gamma(\alpha_0/2)}.\end{aligned}$$

Notice that the last integral converges if $\alpha_0 > 1$. By symmetry of our arguments, we have the same bound for another part of the integral in R_3. Thus,

$$\begin{aligned}|R_3(\alpha)| &\leq \exp(-\overline{\alpha}\kappa - \overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p}, u, f))\frac{\sqrt{\pi}\cdot\Gamma\left(\frac{\alpha_0-1}{2}\right)}{\Gamma(\alpha_0/2)}\left(\sqrt{\frac{(1-\lambda^\star(b_0-u))^2 + K^2(b_0-u)^2}{(b_0-u)^2}}\right) \\ &\leq \frac{\sqrt{5\pi}}{2}\cdot\exp(-c_\kappa\alpha_0 - \overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p}, u, f))\frac{\Gamma\left(\frac{\alpha_0-1}{2}\right)}{\Gamma(\alpha_0/2)}\cdot\frac{1-\lambda^\star(b_0-u)}{b_0-u},\end{aligned}$$

where the last inequality follows from the lower bound $\operatorname{Re}\kappa \geq c_\kappa\alpha_0/\overline{\alpha}$ and the choice of K. Set $c_3 = \sqrt{5\pi}/2$ and note that

$$\frac{\Gamma\left(\frac{\alpha_0-1}{2}\right)}{\Gamma(\alpha_0/2)} \leq 1.$$

Finally, we have

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

□

Our next goal is to provide a lower bound on the density p_Z using the above representation.

Lemma D.13. *Consider a function $f \in \mathrm{F}_m(b_0, b)$ and a vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ with $\overline{\alpha} \geq 2\alpha_0$, $b_0 \geq 2b$ and*

$$\alpha_0 \geq \max\left\{\frac{1}{(\sqrt{2\pi}-1)^2}\cdot\left(\frac{2\sqrt{2}}{\sqrt{\log(17/16)}} + \frac{98\sqrt{6}}{9}\right)^2, \frac{\log(10\pi\cdot\overline{\alpha})}{\log(17/16)}\right\}.$$

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

$$p_Z(u) \geq \frac{\sqrt{\overline{\alpha} - 1/\overline{\alpha}}}{8\pi}\cdot\left(\frac{1-\lambda^\star(b_0-u)}{b_0-u}\right)\cdot\exp(-\overline{\alpha}\,\mathcal{K}_{\mathit{inf}}(\overline{p}, u, f)).$$

Proof. First, we derive from Proposition D.11 a lower bound for the integral (16), using inequality $b_0/(b_0 - \overline{p}f) \leq 4$ under our conditions on b_0 and $\overline{\alpha}$

$$\int_{\mathbb{R}}\prod_{j=0}^{n}(1 + \mathrm{i}(f(j)-u)s)^{-\alpha_j}\mathrm{d}s \geq \frac{\left(\sqrt{2\pi} - \frac{c_1\exp(-c_\kappa\alpha_0)}{\sqrt{c_\kappa\alpha_0}} - \frac{4c_2}{\sqrt{\alpha_0}}\right)}{\sqrt{\overline{\alpha}\,\mathbb{E}[\mathrm{Y}_0^2]}}\exp(-\overline{\alpha}\,\mathcal{K}_{\text{inf}}(\overline{p}, u, f)) + R_3(\alpha),$$

First, we want to ensure that

$$\sqrt{2\pi} - \frac{c_1\exp(-c_\kappa\alpha_0)}{\sqrt{c_\kappa\alpha_0}} - \frac{c_2}{\sqrt{\alpha_0}} \geq 1. \tag{19}$$

To do it, notice that since $\alpha_0 \geq 2$, then $\frac{c_1 \exp(-c_\kappa \alpha_0)}{\sqrt{c_\kappa \alpha_0}} \leq \frac{c_1}{\sqrt{c_\kappa \alpha_0}}$. Therefore, to ensure that (19) holds, we can choose

$$\alpha_0 \geq \frac{1}{(\sqrt{2\pi}-1)^2} \cdot \left(\frac{c_1}{\sqrt{c_\kappa}} + 4c_2 \right)^2 = \frac{1}{(\sqrt{2\pi}-1)^2} \cdot \left(\frac{2\sqrt{2}}{\sqrt{\log(17/16)}} + \frac{98\sqrt{6}}{9} \right)^2.$$

Under these conditions, we derive

$$\begin{aligned}
\int_{\mathbb{R}} \prod_{j=0}^{n} (1 + \mathrm{i}(f(j)-u)s)^{-\alpha_j} \mathrm{d}s &\geq \frac{1}{\sqrt{\overline{\alpha}\, \mathbb{E}[Y_0^2]}} \exp(-\overline{\alpha}\, \mathcal{K}_{\inf}(\overline{p}, u, f)) - |R_3(\alpha)| \\
&\geq \exp(-\overline{\alpha}\, \mathcal{K}_{\inf}(\overline{p}, u, f)) \left(\frac{1}{\sqrt{\overline{\alpha}\, \mathbb{E}[Y_0^2]}} - c_3 \cdot \frac{1 - \lambda^\star (b_0 - u)}{b_0 - u} \exp(-c_\kappa \alpha_0) \right).
\end{aligned}$$

Now using an inequality

$$\mathbb{E}[Y_0^2] \leq \left(\frac{b_0 - u}{2(1 - \lambda^\star(b_0 - u))} + \frac{u}{2(1 + \lambda^\star u)} \right)^2 = \frac{b_0^2}{4(1 - \lambda^\star(b_0 - u))^2 (1 + \lambda^\star u)^2} \leq \frac{4(b_0 - u)^2}{(1 - \lambda^\star(b_0 - u))^2},$$

we conclude that

$$\int_{\mathbb{R}} \prod_{j=0}^{n} (1 + \mathrm{i}(f(j)-u)s)^{-\alpha_j} \mathrm{d}s \geq \frac{1 - \lambda^\star(b_0 - u)}{b_0 - u} \cdot \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\inf}(\overline{p}, u, f))}{\sqrt{\overline{\alpha}}} \left(1/2 - c_3 \exp(-c_\kappa \alpha_0) \cdot \sqrt{\overline{\alpha}} \right).$$

Next, by choosing $\alpha_0 \geq \frac{1}{c_\kappa}(1/2 \cdot \log(\overline{\alpha}) + \log(4c_3)) = \log(10\pi \cdot \overline{\alpha}) / \log(17/16)$, we have

$$\int_{\mathbb{R}} \prod_{j=0}^{n} (1 + \mathrm{i}(f(j)-u)s)^{-\alpha_j} \mathrm{d}s \geq \frac{1 - \lambda^\star(b_0 - u)}{b_0 - u} \cdot \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\inf}(\overline{p}, u, f))}{4\sqrt{\overline{\alpha}}}.$$

Using this bound, we can easily derive a lower bound on the density p_Z

$$\begin{aligned}
p_Z(u) &= \frac{\overline{\alpha} - 1}{2\pi} \int_{\mathbb{R}} \prod_{j=0}^{m} (1 + i(f(j)-u)s)^{-\alpha_j} \mathrm{d}s \\
&\geq \frac{\overline{\alpha} - 1}{2\pi} \cdot \frac{1 - \lambda^\star(b_0 - u)}{b_0 - u} \cdot \frac{\exp(-\overline{\alpha}\, \mathcal{K}_{\inf}(\overline{p}, u, f))}{4\sqrt{\overline{\alpha}}}.
\end{aligned}$$

□

Before we proceed with the proof of Theorem D.2, we need the following auxiliary result.

Lemma D.14. *Consider a function $f \in \mathrm{F}_m(b_0, b)$ and a vector $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_m) \in \mathbb{R}_+^{m+1}$ with $\overline{\alpha} > \alpha_0$. Then for all $\overline{p}f < u < b_0$,*

$$\left(\frac{1}{b_0 - u} - \lambda^\star(\overline{p}, u, f) \right) \geq \frac{\alpha_0}{\overline{\alpha}} \frac{1}{b_0 - u}.$$

Proof. Under the condition $\overline{p}f < u < b_0$, the value $\lambda^\star = \lambda^\star(\overline{p}, u, f)$ satisfies the following equation

$$\mathbb{E}\left[\frac{f(X) - u}{1 - \lambda^\star \cdot (f(X) - u)} \right] = \frac{\alpha_0}{\overline{\alpha}} \frac{b_0 - u}{1 - \lambda^\star \cdot (b_0 - u)} + \sum_{j=1}^{m} \frac{\alpha_j}{\overline{\alpha}} \frac{f(j) - u}{1 - \lambda^\star \cdot (f(j) - u)} = 0. \tag{20}$$

Define a distribution $\widehat{p}$ with $\widehat{p}(i) = \frac{\alpha_i}{\overline{\alpha} - \alpha_0}$ for $i \in \{1, \ldots, m\}$ and $\widehat{p}(0) = 0$. Then the expectation in (20) can be written as

$$\frac{\alpha_0}{\overline{\alpha}} \frac{b_0 - u}{1 - \lambda^\star \cdot (b_0 - u)} + \frac{\overline{\alpha} - \alpha_0}{\overline{\alpha}} \mathbb{E}_{X \sim \widehat{p}}\left[\frac{f(X) - u}{1 - \lambda^\star \cdot (f(X) - u)} \right] = 0.$$

Define a function $w(x,u) = \frac{x-u}{1-\lambda^\star\cdot(x-u)}$, which is convex in x. By the Jensen inequality,

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

Hence

$$\frac{\alpha_0}{\overline{\alpha}}\frac{b_0-u}{1-\lambda^\star\cdot(b_0-u)} \leq -\frac{\overline{\alpha}-\alpha_0}{\overline{\alpha}}\frac{\widehat{p}f-u}{1-\lambda^\star\cdot(\widehat{p}f-u)} = \frac{\overline{\alpha}-\alpha_0}{\overline{\alpha}}\frac{u-\widehat{p}f}{1+\lambda^\star\cdot(u-\widehat{p}f)}.$$

By rearranging terms, we get

$$\frac{1}{\alpha_0}\left(\frac{1}{b_0-u}-\lambda^\star\right) \geq \frac{1}{\overline{\alpha}-\alpha_0}\left(\frac{1}{u-\widehat{p}f}+\lambda^\star\right) = \frac{1}{\overline{\alpha}-\alpha_0}\left(\frac{1}{u-\widehat{p}f}+\frac{1}{b_0-u}\right) - \frac{1}{\overline{\alpha}-\alpha_0}\left(\frac{1}{b_0-u}-\lambda^\star\right).$$

As a result,

$$\left(\frac{1}{b_0-u}-\lambda^\star\right) \geq \frac{\alpha_0}{\overline{\alpha}}\cdot\frac{b_0-\widehat{p}f}{(b_0-u)(u-\widehat{p}f)} \geq \frac{\alpha_0}{\overline{\alpha}}\frac{1}{b_0-u}.$$

$\square$

Now we are ready to finish the proof of Theorem D.2.

Proof of Theorem D.2. By Lemma D.13,

$$\mathbb{P}(Z\geq\mu) = \int_\mu^{b_0} p_Z(u)\mathrm{d}u \geq \frac{\sqrt{\overline{\alpha}-1/\overline{\alpha}}}{8\pi}\cdot\int_\mu^{b_0}\left(\frac{1}{b_0-u}-\lambda^\star(\overline{p},u,f)\right)\cdot\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},u,f))\,\mathrm{d}u.$$

Our goal is to analyze the last integral. First of all, by Lemma D.14 we have

$$\int_\mu^{b_0}\left(\frac{1}{b_0-u}-\lambda^\star(\overline{p},u,f)\right)\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},u,f))\,\mathrm{d}u \geq \frac{\alpha_0}{\overline{\alpha}}\cdot\int_\mu^{b_0}\frac{1}{b_0-u}\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},u,f))\,\mathrm{d}u.$$

Next, by Theorem 6 of Honda and Takemura (2010) we have for all $\overline{p}f<u<b_0$,

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

Therefore, since $\lambda^\star\leq 1/(b_0-u)$, we get

$$\mathcal{K}_{\inf}(\overline{p},u,f)-\mathcal{K}_{\inf}(\overline{p},\mu,f) = \int_\mu^u\lambda^\star(\overline{p},s,f)\mathrm{d}s \leq \int_\mu^u\frac{1}{b_0-s}\mathrm{d}s = \log\left(\frac{b_0-\mu}{b_0-u}\right).$$

Hence,

$$\int_\mu^{b_0}\frac{1}{b_0-u}\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},u,f))\,\mathrm{d}u \geq \exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},\mu,f))\cdot\int_\mu^{b_0}\frac{1}{b_0-u}\exp\left(-\overline{\alpha}\log\left(\frac{b_0-\mu}{b_0-u}\right)\right)\mathrm{d}u,$$

where the last integral could by easily computed

$$\int_\mu^{b_0}\frac{1}{b_0-u}\exp\left(-\overline{\alpha}\log\left(\frac{b_0-\mu}{b_0-u}\right)\right)\mathrm{d}u = \int_\mu^{b_0}\frac{1}{b_0-u}\left(\frac{b_0-u}{b_0-\mu}\right)^{\overline{\alpha}}\mathrm{d}u = \frac{1}{(b_0-\mu)^{\overline{\alpha}}}\int_\mu^{b_0}(b_0-u)^{\overline{\alpha}-1}\mathrm{d}u = \frac{1}{\overline{\alpha}}.$$

Thus,

$$\begin{aligned}
\mathbb{P}(Z\geq\mu) &\geq \frac{\alpha_0}{\overline{\alpha}^2}\cdot\frac{\sqrt{\overline{\alpha}-1/\overline{\alpha}}}{8\pi}\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},\mu,f)) \\
&\geq \frac{\alpha_0\cdot\sqrt{1-\overline{\alpha}^2}}{8\pi}\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},\mu,f)-3/2\log\overline{\alpha}) \\
&\geq \frac{\alpha_0\cdot\sqrt{3}}{16\pi}\exp(-\overline{\alpha}\,\mathcal{K}_{\inf}(\overline{p},\mu,f)-3/2\log\overline{\alpha}),
\end{aligned}$$

where we used that $\overline{\alpha}\geq 2$. To conclude the statement, note that $\alpha_0\geq 16\pi$. $\square$

E. Technical Lemmas

E.1. A Bellman-type equation for the variance

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

$$\begin{aligned}\sigma Q_h^\pi(s,a) &\triangleq \mathrm{Var}_{p_h} V_{h+1}^\pi(s,a) + p_h \sigma V_{h+1}^\pi(s,a)\\ \sigma V_h^\pi(s) &\triangleq \sigma Q_h^\pi(s,\pi(s))\\ \sigma V_{H+1}^\pi(s) &\triangleq 0,\end{aligned}$$

where $\mathrm{Var}_{p_h}(f)(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}\big[(f(s') - p_h f(s,a))^2\big]$ denotes the *variance operator*. In particular, the function $s \mapsto \sigma V_1^\pi(s)$ represents the average sum of the local variances $\mathrm{Var}_{p_h} V_{h+1}^\pi(s,a)$ over a trajectory following the policy π, starting from (s,a). Indeed, the definition above implies that

$$\sigma V_1^\pi(s_1) = \sum_{h=1}^H \sum_{s,a} p_h^\pi(s,a) \mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a).$$

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

Lemma E.1 (Law of total variance)**.** *For any deterministic policy π and for all $h \in [H]$,*

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

In particular,

$$\mathbb{E}_\pi\left[\left(\sum_{h=1}^H r_h(s_h,a_h) - V_1^\pi(s_1)\right)^2\right] = \sigma V_1^\pi(s_1) = \sum_{h=1}^H \sum_{s,a} p_h^\pi(s,a) \mathrm{Var}_{p_h}(V_{h+1}^\pi)(s,a).$$

Proof. We proceed by induction. The statement in Lemma E.1 is trivial for $h = H+1$. We now assume that it holds for $h+1$ and prove that it also holds for h. For this purpose, we compute

$$\begin{aligned}&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^H r_{h'}(s_{h'},a_{h'}) - Q_h^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h)\right]\\ &= \mathbb{E}_\pi\left[\left(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h) + \sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right]\\ &= \mathbb{E}_\pi\left[\left(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h)\right]\\ &+ \mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right]\\ &+ 2\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^H r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)\left(Q_{h+1}^\pi(s_{h+1},a_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\right) \middle| (s_h,a_h)\right].\end{aligned}$$

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

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

Therefore, the law of total expectation gives us

$$
\begin{aligned}
&\mathbb{E}_\pi\left[\left(\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) - Q_h^\pi(s_h,a_h)\right)^2 \middle| (s_h,a_h)\right] \\
&= \mathbb{E}_\pi\Big[\big(V_{h+1}^\pi(s_{h+1}) - p_h V_{h+1}^\pi(s_h,a_h)\big)^2 \Big| (s_h,a_h)\Big] + \mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_h,a_h)\right] \\
&= \mathrm{Var}_{p_h} V_{h+1}^\pi(s_h,a_h) \\
&\quad + \sum_{(s_{h+1},a_{h+1})} p_h(s_{h+1}|s_h,a_h)\mathbb{1}_{(a_{h+1}=\pi(s_{h+1}))}\mathbb{E}_\pi\left[\left(\sum_{h'=h+1}^{H} r_{h'}(s_{h'},a_{h'}) - Q_{h+1}^\pi(s_{h+1},a_{h+1})\right)^2 \middle| (s_{h+1},a_{h+1})\right] \\
&= \mathrm{Var}_{p_h} V_{h+1}^\pi(s_h,a_h) + p_h\sigma V_{h+1}^\pi(s_h,a_h) \\
&= \sigma Q_h^\pi(s_h,a_h)
\end{aligned}
$$

where in the third equality we used the inductive hypothesis and the definition of σV_{h+1}^π. □

E.2. On the Bernstein inequality

We reproduce here a Bernstein-type inequality by Talebi and Maillard (2018).

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

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

Proof. Let $\widetilde{p}$ be the distribution of the pair of random variables (X,Y) where X,Y are i.i.d. according to the distribution p. Similarly, let $\widetilde{q}$ be the distribution of the pair of random variables (X,Y) where X,Y are i.i.d. according to distribution q. Since Kullback–Leibler divergence is additive for independent distributions, we know that

$$
\mathrm{KL}(\widetilde{p},\widetilde{q}) = 2\,\mathrm{KL}(p,q) \le 2\,\mathrm{KL}(p,q).
$$

Using Lemma E.2 for the function $g(x,y) = (f(x)-f(y))^2$ defined on $\mathcal{S}^2$, such that $0 \le g \le b^2$, we get

$$
\begin{aligned}
|\widetilde{p}g - \widetilde{q}g| &\le \sqrt{4\mathrm{Var}_{\widetilde{q}}(g)\,\mathrm{KL}(p,q)} + \frac{4}{3}b^2\,\mathrm{KL}(p,q) \\
&\le \sqrt{4b^2\,\mathrm{KL}(p,q)\widetilde{q}g} + \frac{4}{3}b^2\,\mathrm{KL}(p,q) \\
&\le \frac{1}{2}\widetilde{q}g + \frac{10}{3}b^2\,\mathrm{KL}(p,q)\,,
\end{aligned}
$$

where in the last line we used $2\sqrt{xy} \le x+y$ for $x,y \ge 0$. In particular we obtain

$$
\begin{aligned}
\widetilde{p}g &\le \frac{3}{2}\widetilde{q}g + \frac{10}{3}b^2\,\mathrm{KL}(p,q) \\
\widetilde{q}g &\le 2\widetilde{p}g + \frac{20}{3}b^2\,\mathrm{KL}(p,q)\,.
\end{aligned}
$$

To conclude, it remains to note that

$$\widetilde{p}g = 2\mathrm{Var}_p(f) \text{ and } \widetilde{q}g = 2\mathrm{Var}_q(f).$$

$\square$

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

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

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

Proof. First note that

$$\mathrm{Var}_p(f) = p(f - g + g - pg + pg - pf)^2 \le 2p(f - g - pf + pg)^2 + 2p(g - pg)^2 = 2\mathrm{Var}_p(f - g) + 2\mathrm{Var}_p(g).$$

From the above we can immediately conclude the proof of the first inequality with

$$\mathrm{Var}_p(f - g) \le p(f - g)^2 \le bp|f - g|,$$

where we used that for all $s \in \mathcal{S}$, $0 \le |f(s) - g(s)| \le b$. For the second inequality, using the Hölder inequality,

$$\begin{aligned}
\mathrm{Var}_q(f) &= pf^2 - (pf)^2 + (q - p)f^2 + (pf)^2 - (qf)^2 \\
&\le \mathrm{Var}_p(f) + b^2\|p - q\|_1 + 2b^2\|p - q\|_1 \\
&\le \mathrm{Var}_p(f) + 3b^2\|p - q\|_1.
\end{aligned}$$

$\square$

E.3. Inequalities for quantiles

Lemma E.5. *Let X be a random variable from distribution ρ, $\kappa \in (0, 1)$, and $u \in \mathbb{R}$. Then*

$$\mathbb{P}_{X\sim p}(X > u) \le 1 - \kappa \iff \mathbb{Q}_{X\sim p}(X, \kappa) \le u, \qquad \mathbb{P}_{X\sim p}(X > u) \ge 1 - \kappa \iff \mathbb{Q}_{X\sim p}(X, \kappa) \ge u.$$

Proof. Follows directly from the definition

$$\mathbb{Q}_{X\sim p}(X, \kappa) = \inf\{t \in \mathbb{R} : \mathbb{P}_{X\sim p}(X > u) \le 1 - \kappa\}.$$

$\square$

E.4. Jacobian computation

For any non-zero vector $v \in \mathbb{R}^n$ we define linear parametrization of the set $\mathrm{H}_v^t = \{x \in \mathbb{R}^n | v^\top x = t\}$ by a linear map $\mathcal{L}_v^t \colon \mathbb{R}^{n-1} \to \mathbb{R}^n$ by a rule $y \to x, y_1 = x_1, \dots, y_{n-1} = x_{n-1}, x_0 = \frac{t - \sum_{i=1}^{n-1} v_i y_i}{v_0}$. Without loss of generality we may assume that $v_0 \neq 0$ so the parametrization is well-defined. Using matrix language, we can define a matrix $\mathcal{L}_v \in \mathbb{R}^{n \times n-1}$ with the first row is equal to $[-v_1/v_0, \dots, v_m/v_0]$, and the last $n - 1$ row forms an identity matrix. For a linear map $\mathcal{L} \colon \mathbb{R}^n \to \mathbb{R}^m$ we define a Jacobian $[\mathcal{L}]$ as a generalized determinant of a gradient matrix of this map (see (Evans and Garzepy, 2018) for more detail).

Lemma E.6 (Jacobian of a linear parametrization)**.** *For any non-zero vector $v \in \mathbb{R}^n$ such that $v_0 \neq 0$ and $t \in \mathbb{R}$ a Jacobian of map $\mathcal{L}_v^t$ is equal to $\|v\|_2/|v_0|$.*

Proof. Note that the gradient vector does not depend on constant shifts. Thus the gradient matrix is equal to a linear map $\mathcal{L}_v$. Define a vector $\tilde{v} = [v_1, \dots, v_n]$, then the square Jacobian is equal to

$$[\mathcal{L}_v]^2 = \det(\mathcal{L}_v^\top \mathcal{L}_v) = \det\left(\begin{bmatrix} \tilde{v}/v_0 & I_{n-1} \end{bmatrix} \begin{bmatrix} \tilde{v}^\top/v_0 \\ I_{n-1} \end{bmatrix}\right) = \det\left(\frac{\tilde{v}\tilde{v}^\top}{v_0^2} + I_{n-1}\right).$$

This matrix is a rank-one matrix plus identity. Its eigenvalues are equal to $\|\tilde{v}\|^2/v_0^2 + 1$ and $n - 2$ ones. Thus $[\mathcal{L}_v]^2 = \|\tilde{v}\|^2/v_0^2 + 1 = \|v\|^2/v_0^2$. $\square$

Lemma E.7 (Jacobian of a composition of linear parametrizations)**.** *For a vector $f \in \mathbb{R}^{m+1}$ define a vector $c = \mathcal{L}_{\mathbf{1}}^{\intercal} f$. Assume that $f_0 \neq f_1$. Then*

$$[\mathcal{L}_{\mathbf{1}}\mathcal{L}_c]^2 = \frac{(m+1)\sum_{j=0}^m f_j^2 - \left(\sum_{j=0}^m f_j\right)^2}{(f_1 - f_0)^2}.$$

Proof. To compute a Jacobian of the composition we have to compute a vector c, a matrix $\mathcal{L}_c$ and a matrix $\mathcal{L}_{\mathbf{1}}\mathcal{L}_c$.

$$\mathcal{L}_{\mathbf{1}} = \begin{bmatrix} -1 & -1 & \dots & -1 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}, \qquad c^{\intercal} = \mathcal{L}_{\mathbf{1}}^{\intercal} f = \begin{bmatrix} f_1 - f_0 \\ \vdots \\ f_m - f_0 \end{bmatrix}, \qquad \mathcal{L}_c = \begin{bmatrix} \frac{f_0 - f_2}{f_1 - f_0} & \frac{f_0 - f_3}{f_1 - f_0} & \dots & \frac{f_0 - f_m}{f_1 - f_0} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

Finally,

$$A = \mathcal{L}_{\mathbf{1}}\mathcal{L}_c = \begin{bmatrix} \frac{f_2 - f_1}{f_1 - f_0} & \frac{f_3 - f_1}{f_1 - f_0} & \dots & \frac{f_m - f_1}{f_1 - f_0} \\ \frac{f_0 - f_2}{f_1 - f_0} & \frac{f_0 - f_3}{f_1 - f_0} & \dots & \frac{f_0 - f_m}{f_1 - f_0} \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Call $v = \left(\frac{f_2 - f_1}{f_1 - f_0}, \frac{f_3 - f_1}{f_1 - f_0}, \dots, \frac{f_m - f_1}{f_1 - f_0}\right)^{\intercal}$ and $u = \left(\frac{f_0 - f_2}{f_1 - f_0}, \frac{f_0 - f_3}{f_1 - f_0}, \dots, \frac{f_0 - f_m}{f_1 - f_0}\right)^{\intercal}$ the first and the second rows of the matrix A. We have to compute the following determinant

$$\det(A^{\intercal}A) = \det\left(\begin{bmatrix} v & u & I \end{bmatrix}\begin{bmatrix} v^{\intercal} \\ u^{\intercal} \\ I \end{bmatrix}\right) = vv^{\intercal} + uu^{\intercal} + I.$$

Notice that this matrix is a rank-2 matrix plus the identity matrix. Let λ_1, λ_2 be two non-zero eigenvalues. Then eigenvalues of matrix $A^{\intercal}A$ is $\lambda_1 + 1, \lambda_2 + 1$ and all other eigenvalues are ones. Thus

$$\det(A^{\intercal}A) = (\lambda_1 + 1)(\lambda_2 + 1) = \lambda_1\lambda_2 + (\lambda_1 + \lambda_2) + 1.$$

To compute it, notice that non-zero eigenvalues of $B^{\intercal}B$ and $BB^{\intercal}$ coincide. Then we note

$$vv^{\intercal} + uu^{\intercal} = \begin{bmatrix} v & u \end{bmatrix}\begin{bmatrix} v^{\intercal} \\ u^{\intercal} \end{bmatrix},$$

therefore λ_1, λ_2 are eigenvalues of the following matrix 2×2

$$\begin{bmatrix} v^{\intercal} \\ u^{\intercal} \end{bmatrix}\begin{bmatrix} v & u \end{bmatrix} = \begin{bmatrix} \|v\|^2 & v^{\intercal}u \\ u^{\intercal}v & \|u\|^2 \end{bmatrix} = C$$

Then we note that $\det(C) = \lambda_1\lambda_2$ and $\lambda_1 + \lambda_2 = \mathrm{Tr}(C)$. Thus

$$\det(A^{\intercal}A) = \|v\|^2\|u\|^2 - (v^{\intercal}u)^2 + \|v\|^2 + \|u\|^2 + 1.$$

Next we start computing required quantities

$$\|v\|^2 = \frac{\sum_{i=2}^m (f_1 - f_i)^2}{(f_1 - f_0)^2}, \qquad \|u\|^2 = \frac{\sum_{i=2}^m (f_0 - f_i)^2}{(f_1 - f_0)^2}, \qquad |v^{\intercal}u| = \frac{\sum_{i=2}^m (f_1 - f_i)(f_0 - f_i)}{(f_1 - f_0)^2},$$

and then

$$
\begin{aligned}
\|v\|^2\|u\|^2-(v^\intercal u)^2 &= \frac{1}{(f_1-f_0)^4}\left(\sum_{i,j=2}^{m}(f_1-f_i)^2(f_0-f_j)^2-\sum_{i,j=2}^{m}(f_1-f_i)(f_1-f_j)(f_0-f_i)(f_1-f_j)\right)\\
&= \frac{1}{(f_1-f_0)^4}\left(\sum_{i,j=2}^{m}(f_1-f_i)(f_0-f_j)[(f_1-f_i)(f_0-f_j)-(f_1-f_j)(f_0-f_i)]\right)\\
&= \frac{1}{(f_1-f_0)^3}\left(\sum_{i,j=2}^{m}(f_1-f_i)(f_0-f_j)(f_i-f_j)\right)\\
&= \frac{1}{(f_1-f_0)^3}\left(\sum_{i,j=2}^{m}(f_0f_1-f_0f_i-f_1f_j+f_if_j)(f_i-f_j)\right)\\
&= \frac{1}{(f_1-f_0)^3}\sum_{i,j=2}^{m}(f_0f_1f_i-f_0f_i^2-f_1f_if_j+f_i^2f_j-f_0f_1f_j+f_0f_if_j+f_1f_j^2-f_if_j^2).
\end{aligned}
$$

Define $S=\sum_{i=2}^m f_i, V=\sum_{i=2}^m f_i^2$. Then after grouping the terms

$$
\begin{aligned}
\|v\|^2\|u\|^2-(v^\intercal u)^2 &= \frac{f_0f_1(m-1)S-f_0(m-1)V-f_1S^2+VS-f_0f_1(m-1)S+f_0S^2+f_1(m-1)V-SV}{(f_1-f_0)^3}\\
&= \frac{(m-2)V-S^2}{(f_1-f_0)^2}.
\end{aligned}
$$

Finally

$$
\begin{aligned}
\det(A^\intercal A) = \frac{1}{(f_1-f_0)^2}\Bigg((m-1)\sum_{i=2}^{m}f_i^2-\left(\sum_{i=2}^{m}f_i\right)^2+(m-1)f_0^2-2f_0\sum_{i=2}^{m}f_i+\sum_{i=2}^{m}f_i^2\\
+(m-1)f_1^2-2f_1\sum_{i=2}^{m}f_i+\sum_{i=2}^{m}f_i^2+2f_0^2+2f_1^2-2f_0f_1-f_0^2-f_1^2\Bigg)\\
= \frac{(m+1)\sum_{i=0}^m f_i^2-\left(\sum_{i=0}^m f_i\right)^2}{(f_1-f_0)^2}.
\end{aligned}
$$

□

F. Non-tabular Extension Detailed

In this section, we detail the extension of Bayes-UCBVI beyond the tabular setting. First recall that in Incr-Bayes-UCBVI, Q-value functions are given by the quantile $\overline{Q}_h^t(s,a) = \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}([B])}\Big(\overline{Q}_h^{t,b}(s,a), \kappa_h^t(s,a)\Big)$ over B (incremental) Bayesian bootstrap samples. Theses samples can be computed by updating sums of random weights distributed according to an exponential distribution of parameter one. Precisely, we define

$$Z_h^{t,b}(s,a,s') \triangleq \sum_{\ell=1}^{t} \mathbb{1}\{(s_h^\ell, a_h^\ell) = (s,a)\} z_h^{\ell,b}(s,a) + \mathbb{1}\{s' = s_0\} \sum_{\ell=-n_0+1}^{0} z_h^{\ell,b}(s,a)\,,$$

where the weights are i.i.d. $z_h^{t,b}(s,a) \sim \mathcal{E}(1)$. Then, the Bayesian bootstrap sample at time t is given by

$$\overline{Q}_h^{t,b}(s,a) = r_h(s,a) + \sum_{s'\in\mathcal{S}'} \frac{Z_h^{t,b}(s,a,s')}{\sum_{s''\in\mathcal{S}'} Z_h^{t,b}(s,a,s'')} \overline{V}_{h+1}^t(s')\,.$$

The complete procedure is detailed in Algorithm 2. Note that if the rewards are unknown we just need to also re-weight the observed rewards for a given state-action pair, similarly to Riou and Honda (2020) for multi-armed bandits. There are two approximations made in Incr-Bayes-UCBVI with respect to Bayes-UCBVI. First, Incr-Bayes-UCBVI approximates the quantile with B Monte-Carlo samples. Second, Incr-Bayes-UCBVI reuses the same exponential noise from one time step to the next one to improve the time complexity (run-time).

Algorithm 2 Incr-Bayes-UCBVI

1: **Input:** quantile κ, B number of bootstrap samples, number of prior transitions n_0.
2: **Initialize:** weights $Z_h^0(s,a,s') \leftarrow \mathbb{1}\{s' = s_0\}\sum_{\ell=-n_0+1}^{0} z_h^{\ell,b}(s,a)$, where $z_h^{\ell,b} \sim \mathcal{E}(1)$ i.i.d., for $h,s,a,s' \in [H]\times\mathcal{S}\times\mathcal{A}\times\mathcal{S}'$.
3: **for** $t \in [T]$ **do**
4: Optimistic value iteration

$$\overline{Q}_h^{t-1}(s,a) \leftarrow \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}[B]}\left(r_h(s,a) + \sum_{s'\in\mathcal{S}'} \frac{Z_h^{t-1,b}(s,a,s')}{\sum_{s''\in\mathcal{S}'} Z_h^{t-1,b}(s,a,s'')} \overline{V}_{h+1}^{t-1}(s'), \kappa_h^{t-1}(s,a)\right)$$
$$\overline{V}_h^{t-1}(s) \leftarrow \max_{a\in\mathcal{A}} \overline{Q}_h^{t-1}(s,a) \qquad \overline{V}_h^{t-1}(s_0) \leftarrow V_h^\star(s_0)$$
$$\overline{V}_{H+1}^{t-1}(s) \leftarrow 0\,.$$

5: **for** $h \in [H]$ **do**
6: Play $a_h^t \in \arg\max_a \overline{Q}_h^{t-1}(s,a)$.
7: Observe reward r_h^t and next state $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$.
8: Update weights $Z_h^{t,b}(s,a,s') \leftarrow Z_h^{t-1,b}(s,a,s') + \begin{cases} z_h^{t,b}(s,a,s') \sim \mathcal{E}(1) \text{ i.i.d.} & \text{if } (s,a,s') = (s_h^t, a_h^t, s_{h+1}^t) \\ 0 & \text{else} \end{cases}$.
9: **end for**
10: **end for**

We now present the non-tabular extension of Incr-Bayes-UCBVI. Fix a state-action pair (s,a) at time t. As explained in Section 4, the Bayesian bootstrap samples can be obtained thanks to a weighted linear regression,

$$\overline{Q}_h^{t,b}(s,a) \triangleq \arg\min_x \sum_{n=-n_0+1}^{n_h^t(s,a)} z_h^{n,b}(s,a)(x - y_h^n(s,a))^2, \qquad \text{where } z_h^{n,b}(s,a) \sim \mathcal{E}(1) \text{ i.i.d.}$$

and the targets are defined by $y_h^n(s,a) \triangleq r_h(s,a) + \overline{V}_{h+1}^t(s_{h+1}^n)$. Note that it is possible to adapt this re-weighting to any loss instead of the mean-squared error. To combine this exploration procedure with the DQN algorithm, we introduce two Q-value networks. One Q-value network Q^ψ parametrized by ϕ and one target Q-value network parametrized by ψ. We also

Algorithm 3 Bayes-UCBDQN

Input: discount factor γ, quantile κ, B number of bootstrap samples, ϕ, ψ parameter of the Q-value respectively target Q-value network, replay buffer $\mathcal{R}$, pseudo target $y^{\texttt{pseudo}}$, pseudo transition probability ε.
for $t \in [T]$ **do**
  for $h \in [H]$ **do**
    Play $a_h^t \in \arg\max_a \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}[B]}\big(Q_h^{\phi,b}(s_h^t, a), \kappa\big)$.
    Observe reward r_h^t and next state $s_{h+1}^t \sim p_h(s_h^t, a_h^t)$.
    Record $\{h, r_h^t, s_h^t, a_h^t, s_{h+1}^t, z_h^t\}$ in $\mathcal{R}$ where $z_h^{t,b} \sim \mathcal{E}(1)$ i.i.d for $b \in [B]$.
    With probability ε record $\{h, \_, s_h^t, a^0, s_0, z^0\}$ in $\mathcal{R}$ where $z^{0,b} \sim \mathcal{E}(1)$ i.i.d for $b \in [B]$ and $a^0 \sim \mathcal{U}\mathrm{nif}(\mathcal{A})$.
  end for
  if time to update **then**
    Sample a batch of transitions $\mathcal{C} = \big\{(h, r, s, a, s', z)\big\}$ from $\mathcal{R}$
    Compute the targets
    for $(h, r, s, a, s', z) \in \mathcal{C}$ **do**
      if $s \neq s_0$ **then**

$$y(h,s,a,s') \leftarrow r + \gamma \max_{a'\in\mathcal{A}} \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}[B]}\big(Q_{h+1}^{\psi,b}(s', a'), \kappa\big)$$

      else

$$y(h,s,a,s') \leftarrow y^{\texttt{pseudo}} .$$

      end if
    end for
    Update the Q-value network by one step of gradient descent with

$$\nabla_\phi \frac{1}{|\mathcal{C}|} \sum_{(h,s,a,s',z)\in\mathcal{C}} \sum_{b=1}^{B} z^b \Big(Q_h^{\phi,b}(s,a) - y(h,s,a,s')\Big)^2 .$$

    if time to update target **then**
      Update the target Q-value network $\psi \leftarrow \phi$.
    end if
  end if
end for

introduce a replay buffer $\mathcal{R}$. For each observed transition, we record in $\mathcal{R}$ the tuple $\{h, r_h^t, s_h^t, a_h^t, s_{h+1}^t, z_h^t\}$, where we also add an exponential mask z_h^t. Each element of the mask $z_h^{t,b} \sim \mathcal{E}(1)$ i.i.d. , will be used to weight the loss. Precisely, given a batch of transitions $\mathcal{C} = \big\{(h, r, s, a, s', z)\big\}$; from $\mathcal{R}$ we first compute the targets with the target Q-value network,

$$y(h,s,a,s') \leftarrow r + \gamma \max_{a'\in\mathcal{A}} \mathbb{Q}_{b\sim\mathcal{U}\mathrm{nif}[B]}\big(Q_{h+1}^{\psi,b}(s', a'), \kappa\big) .$$

Then, we update each (incremental) Bayesian bootstrap of the Q-value network with one gradient step

$$\nabla_\phi \frac{1}{|\mathcal{C}|} \sum_{(h,s,a,s',z)\in\mathcal{C}} \sum_{b=1}^{B} z^b \Big(Q_h^{\phi,b}(s,a) - y(h,s,a,s')\Big)^2 .$$

As in DQN, the target Q-value network is regularly updated with the weights of the Q-value network. To stick to the Incr-Bayes-UCBVI, it remains to introduce a mechanism that emulates the prior transitions. Since obviously, we cannot add a prior transition for each state-action pair, we propose to simply use ε-greedy *to add prior transitions* to the replay buffer $\mathcal{R}$. In particular, when interacting with the environment, say that the agent is in state s_h^t, then with probability ε, we add a pseudo transition $\{h, \_, s_h^t, a_0, s_0, z_h^0\}$ to the replay buffer $\mathcal{R}$, where the action $a_0 \sim \mathcal{U}\mathrm{nif}(\mathcal{A})$ is sampled uniformly at random and z_h^0 is an exponential mask. When a pseudo transition is sampled from the replay buffer, we assign to it an arbitrary fixed target $y^{\texttt{pseudo}}$. Typically, we want the constant $y^{\texttt{pseudo}}$ to be large, e.g., of the order of the value of the pseudo state s_0. We name the resulting algorithm `Bayes-UCBDQN` and detail it in Algorithm 3. We highlight in blue the procedure used to add the pseudo transitions.

G. Experimental Details

In this section we detail the experiments we conduct for tabular and large-scale environments.

G.1. Tabular

Environment For the tabular experiments[10] we consider a simple grid-world with 5 connected rooms of size 5×5 totalling $S = 129$ states. The agents starts in the middle where there is a very small deterministic reward of 0.01. Additionally there is one small deterministic reward of 0.1 in the leftmost room, one large deterministic reward of 1 in the rightmost room and zero reward elsewhere. The agent can take $A = 4$ actions: moving up, down, left, right. When taking an action, the agent moves in the corresponding direction with probability 0.9 and moves to a neighbor state at random with probability 0.1. The horizon is fixed to $H = 30$. Thus in this environment, the agent must explore efficiently all the rooms avoiding being lured by the small reward in the leftmost room.

Baselines We compare `Bayes-UCBVI` and `Incr-Bayes-UCBVI` to the following baselines:

- `PSRL` (Osband et al., 2013),
- `RLSVI` (Osband et al., 2016b),
- `UCBVI` (Azar et al., 2017).

Using different parameters (e.g., by changing the multiplicative constants in the bonus of `UCBVI` or the scale of the noise in `RLSVI`) can result in drastically different empirical regrets. Thus for a fair comparison, we make the following choices. For `UCBVI`, the bonus at state-action pair is

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

As explained by Ménard et al. (2021), this bonus does not necessarily result in a true UCB on the optimal Q-value. However, it is a valid UCB for $n_h^t(s,a) = 0$ which is important in order to discover new state-action pairs. For `RLSVI` at state-action pair (s,a), we use noise from a centered Gaussian probability distribution with standard deviation $\beta_h^t(s,a)$. For `PSRL`, we use a Dirichlet prior on the transition probability distribution with parameter $(1/S, \ldots, 1/S)$ and for the rewards a Beta prior with parameter $(1,1)$. Note that since the reward r is not necessarily in $\{0,1\}$ we just sample a new reward $r' \sim \mathcal{B}\mathrm{er}(r)$ accordingly to a Bernoulli distribution of parameter r, to update the posterior, see Garivier and Cappé (2011). We choose the same parameter for `Bayes-UCBVI` and `Incr-Bayes-UCBVI`. The quantile is fixed to $\kappa_h^t(s,a) \triangleq 0.85$. This choice is informally justified as follows: If we assume $\beta_h^t(s,a) \approx \sqrt{\big(H^2\log(1/\delta)\big)/\big(2n_h^t(s,a)\big)}$ then it holds $\delta \approx e^{-2/H^2} \geq e^{-2} \approx 0.15$. Thus, we get $\kappa \approx 1-\delta \approx 0.85$. The number of pseudo transitions is set to $n_0 = 1$ as well as the reward of the pseudo transition is $r_0 = 1$. We use $B = 64$ Monte-Carlo samples to approximate the quantile.

Results In Figure 1, we plot the regret of the various baselines, `Bayes-UCBVI` and `Incr-Bayes-UCBVI` in the aforementioned environment. In this experiment, we observe that both `Bayes-UCBVI` and `Incr-Bayes-UCBVI` achieves competitive results with respect to baselines relying on noise-injection for exploration (`PSRL`, `RLSVI`). This is remarkable, since the common belief is that optimistic algorithm perform poorly in practice (Osband and Van Roy, 2017). Indeed `Incr-Bayes-UCBVI` exhibits a similar regret than `PSRL`. It is not completely surprising since they share the same model on the transitions[11] (up to the prior). Whereas `Bayes-UCBVI` performs slightly worse than `Incr-Bayes-UCBVI` but better than `RLSVI`. One possible reason to explain this gap between `Bayes-UCBVI` and `Incr-Bayes-UCBVI` is that the incremental implementation of Bayesian bootstrap forgets faster the prior than the non-incremental version resulting in a more aggressive algorithm. We also note that `RLSVI` performs slightly worse than `PSRL`/`Incr-Bayes-UCBVI` but much better than `UCBVI`. A possible explanation for this ranking is that `RLSVI` is much more aggressive than `UCBVI` when they have comparable noise/bonuses. Whereas `PSRL`,`Incr-Bayes-UCBVI`,`Bayes-UCBVI` take better advantage of the small variance of this particular environment than the two last baselines.

[10]Our code is based on the library of Domingues et al. (2021a).

[11]However, not exactly the same on the rewards: Beta/Bernoulli versus non-parametric. Nonetheless, this should have a relatively small impact since the rewards are deterministic.

G.2. Deep RL

We introduce important details of a few baseline algorithms.

`DQN`. `DQN` adopts the standard Q-learning algorithm. `DQN` maintains a Q-function Q_θ. When interacting with the environment, `DQN` adopts ε-greedy with respect to $Q_\theta(x, a)$. The value of ε decays linearly from $\varepsilon_{\text{max}} = 1.0$ to $\varepsilon_{\text{min}} = 0.01$. The algorithm puts transition tuples (s, a, r, s') into a replay buffer $\mathcal{R}$. At training time, `DQN` samples $C = 32$ transitions tuples $(s_i, a_i, r_i, s'_i)_{i=1}^C \sim \mathcal{R}$ uniformly from the replay. It updates the parameter by minimizing the following loss function with respect to θ,

$$\frac{1}{C}\sum_{i=1}^{C}\Big(Q_\theta(s_i, a_i) - r_i - \gamma \max_{a'} Q_{\theta^-}(s'_i, a')\Big)^2.$$

Here θ^- is the parameter of the target network, a delayed copy of θ used for stabilizing training (Mnih et al., 2013). See (Mnih et al., 2013) for other missing hyper-parameters.

Double `DQN`. Double `DQN` (Van Hasselt et al., 2016), built on top of `DQN` (Mnih et al., 2013), further applies the double Q-learning algorithm to stabilize learning (Hasselt, 2010). Concretely, in addition to the online network θ, Double `DQN` maintains two copies of online networks θ_1, θ_2; and two copies of target networks θ'_1, θ'_2. Double `DQN` constructs the following learning target:

$$Q_1^{(i)} = r_i + \gamma Q_{\theta_1^-}(s'_i, a'_i), a'_i = \arg\max_{a'} Q_{\theta_2^-}(s_i, a')$$
$$Q_2^{(i)} = r_i + \gamma Q_{\theta_2^-}(s'_i, a'_i), a'_i = \arg\max_{a'} Q_{\theta_1^-}(s_i, a').$$

Then both networks θ_1, θ_2 are updated by minimizing the following loss function

$$\frac{1}{C}\sum_{i=1}^{C}\Big(Q_{\theta_1}(s_i, a_i) - Q_1^{(i)}\Big)^2 + \frac{1}{C}\sum_{i=1}^{C}\Big(Q_{\theta_2}(s_i, a_i) - Q_2^{(i)}\Big)^2.$$

In summary, Double `DQN` decouples the maximizing operation and the maximizing value. Empirically, this helps avoid the over-estimation issue that plagues vanilla `DQN`. Similar techniques have been implemented in `BootDQN` and we adopt the same technique in `Bayes-UCBDQN`.

`BootDQN`. `BootDQN` (Osband and Van Roy, 2015) maintains B copies of the Q-function $Q_{\theta_b}, b \in [B]$. In practice, maintaining B full networks might be too expensive; instead, Osband and Van Roy (2015) suggested to share the torso network and only maintain $B = 10$ different heads to the Q-networks. During interaction with the environment, `BootDQN` uniformly random samples a network $Q_{\theta_b}, b \sim \text{Uniform}[B]$. Then the algorithm selects an action by being greedy with respect to Q_{θ_b}. In practice, it has been observed that some local greedy exploration might also be helpful (Osband and Van Roy, 2015; Osband et al., 2018). The transition (s, a, r, s') is put into the buffer $\mathcal{R}$, along with a mask $m \in \mathbb{R}^B$ with each component independently generated from a Bernoulli distribution of parameter $p \in [0, 1]$. At training time, only the Q-networks with mask $m_b = 1$ is trained with the sampled transition using the `DQN` loss function.

Though the mask is meant to enforce diversity across different copies of the Q-networks. In practice, Osband and Van Roy (2015) reported that when $p = 1$ the method works well too. They speculated it is because the stochastic gradient based training of `DQN` networks already led to sufficient amount of diversity.

`Bayes-UCBDQN`. Similar to `BootDQN`, `Bayes-UCBDQN` maintains $B = 10$ copies of the Q-functions by creating B separate Q-network heads with a shared torso network. At acting time, `Bayes-UCBDQN` acts greedily with respect to the bootstrap quantile, computed across B heads. The transition (s, a, r, s') is put into the buffer $\mathcal{R}$, along with a mask $z \sim \mathcal{E}(1)$. At training time, this mask is used for weighing different transitions (see Algorithm 3). All Q-networks are updated by a common target, computed as the κ-quantile bootstrapped values. The quantile is set at $\kappa = 0.85$ in our experiments.

We provide a detailed discussion on the effect of different hyper-parameters and implementations on the performance of the agent, such as the number of heads B and quantile parameter κ. See Appendix G.3.

Network architecture. The network architecture follows from Mnih et al. (2013). The network consists of a torso network with convolutional layers that process the input state images s. The torso layer outputs an embedding $\text{embed} = \text{torso}(s)$. The downstream network is a MLP that takes the embedding as input, and output a A-dimensional vector as the Q-function approximation $\text{head}(\text{embed}) \in \mathbb{R}^A$; see Mnih et al. (2013) for detailed definitions of layer sizes and non-linear activation functions.

For `BootDQN`, in order to avoid the high computational complexity of having B copies of the full network, they argued to instead maintain H copies of the head networks $\text{head}_b, b \in [B]$. The Q-function for the b-th head is defined as $Q_{\theta_b}(s, a) = \text{head}_b(\text{embed})$.

Other hyper-parameters. All algorithms are trained with 200M frames for each environment. The networks are trained with RMSProp optimizer (Hinton et al., 2012) with learning rate $2.5 \cdot 10^{-4}$. See Mnih et al. (2013) and Osband and Van Roy (2015) for other missing hyper-parameters.

Environment and evaluation. The testing environments are Atari-57 games, consisting of 57 selected Atari games (Bellemare et al., 2013). For each game, the state s_t is an image and the action a_t corresponds to controls in the game. For each game, at iteration t, let $z_t^{(i)}$ be the performance of the algorithm in a particular game i for $1 \leq i \leq 57$. Then the human-normalized performance is normalized per game as

$$\text{norm}(z_t^{(i)}) = \frac{z_t^{(i)} - z_u^{(i)}}{z_h^{(i)} - z_u^{(i)}}.$$

Here, $z_u^{(i)}$ is the performance of the random policy in game i; $z_h^{(i)}$ is the human performance in game i. The median performance at iteration t across all game is computed as

$$\text{median-performance}(t) = \text{median}\left(\{z_t^{(t)}\}_{i=1}^{57}\right).$$

The operation $\text{median}(\cdot)$ takes a set of human normalized scores per game and computes the median value over all scores.

G.3. Discussion on the effect of hyper-parameters on `Bayes-UCBDQN`

As shown in Algorithm 3, `Bayes-UCBDQN` adds a considerable number of hyper-parameters compared to `DQN` and `BootDQN`. In our experiments, we have empirically studied the effect of such hyper-parameters and provide a detailed discussion here.

Effect of κ. Throughout experiments, we use $\kappa = 0.85$. We have tested the algorithm's performance under other values of $\kappa \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$. Overall, we find that the algorithm's performance is not very sensitive to κ.

Pseudo transition probability ε and pseudo target y^{pseudo}. Throughout experiments, we use $\varepsilon = 0$. This is mainly because, we find that the training tend to be unstable when using ε significantly larger than 0. The best choice of y^{pseudo} seems to be also game-dependent, and as a result, is more challenging to tune in practice.

We speculate that the instability is due to the fact that `DQN` updates are based on the minimization of squared Bellman errors. Concretely, when with probability ε, the algorithm encounters y^{pseudo}, which is an optimistic value estimate and is hence likely to be an outlier in the data distribution over target values, the update becomes unstable. To address such issues might require further modification to the `DQN` updates, such as based on categorical representation of values (Schrittwieser et al., 2020) or alternative update rules (Bas-Serrano et al., 2021).

Exponential mask z^b. Throughout experiments, we sample $z \sim \mathcal{E}(1)$ for each transition and apply the mask during learning, based on Algorithm 3. Empirically, we find that the mask plays a similar role as the Bernoulli masks adopted in (Osband and Van Roy, 2015).

Effect of number of heads B. Throughout experiments, we use $B = 10$. We have also tried $B \in \{10, 20, 100\}$. Overall, we find that increasing the number of heads does not improve the performance. Quite on the contrary, larger number of heads slightly degrades the performance. Such empirical ablations are consistent with the observation of Osband and Van Roy (2015) that $B \approx 10$ works the best.