Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Why might Posterior Sampling not be better than Optimism in Reinforcement Learning?

Jian Qian JIAN.QIAN@ENS.FR
Sequel Team - Inria Lille

Ronan Fruit RONAN.FRUIT@INRIA.FR
Sequel Team - Inria Lille

Matteo Pirotta MATTEO.PIROTTA@INRIA.FR
Sequel Team - Inria Lille

Alessandro Lazaric LAZARIC@FB.COM
Facebook AI Research

Editor: Kevin Murphy and Bernhard Schölkopf

Abstract

Posterior Sampling (PS) is an effective method to solve the exploration-exploitation dilemma in a wide range of problems. Furthermore, empirical evidence shows that it often outperforms alternative approaches based on the optimism-in-face-of-uncertainty (OFU) principle. In reinforcement learning, this superiority has been supported by theoretical results showing that the regret of PS is smaller than equivalent OFU algorithms. In particular, consider a communicating MDP with S states, A actions, $\Gamma \leq S$ possible next states at each transition, and diameter D. While UCRL (Jaksch et al., 2010) (based on the OFU principle) achieves a regret bound of $\widetilde{O}(D\sqrt{\Gamma SAT})$, an optimistic version of PS (Agrawal and Jia, 2017) reduces the regret to $\widetilde{O}(D\sqrt{SAT})$, thus saving a $\sqrt{\Gamma}$ factor ("pure" PS achieves a similar bound for the Bayesian regret (Osband and Roy, 2017)). In this paper, we investigate the main technical tool used to achieve such improvement and show that it is flawed, thus raising the question on whether PS does indeed improve over UCRL or not.

Keywords: Reinforcement Learning, Posterior Sampling, Optimism, Exploration-Exploitation

1. Introduction

Reinforcement learning (RL) (Sutton and Barto, 1998) studies the problem of learning in sequential decision-making problems where the dynamics of the environment is unknown, but can be learnt by performing actions and observing their outcome in an online fashion. A sample-efficient RL agent must trade off the *exploration* needed to collect information about the environment, and the *exploitation* of the experience gathered so far to gain as much reward as possible. In this paper, we focus on the regret framework in *infinite-horizon average-reward* problems (Jaksch et al., 2010), where the exploration-exploitation performance is evaluated by comparing the rewards accumulated by the learning agent and an optimal policy. Jaksch et al. (2010) showed that it is possible to efficiently solve the exploration-exploitation dilemma using the *optimism in face of uncertainty* (OFU) principle. OFU methods build confidence intervals on the dynamics and reward (i.e., construct a set of plausible MDPs), and execute the optimal policy of the "best" MDP in the confidence region (e.g., Jaksch

et al., 2010; Bartlett and Tewari, 2009; Fruit et al., 2017; Talebi and Maillard, 2018; Fruit et al., 2018). An alternative approach is posterior sampling (PS) (Thompson, 1933), which maintains a posterior distribution over MDPs and, at each step, samples an MDP and executes the corresponding optimal policy (e.g., Osband et al., 2013; Abbasi-Yadkori and Szepesvári, 2015; Osband and Roy, 2017; Ouyang et al., 2017; Agrawal and Jia, 2017). Both OFU and PS approaches rely on estimates of the unknown MDP and the way the estimated MDP concentrates to the true one plays a central role in analysing the theoretical performance of the algorithms.

Given a finite MDP with S states, A actions, and diameter D (i.e., the longest shortest path between any two states), Jaksch et al. (2010) proved that, in the settings of *infinite-horizon average-reward* problems, no algorithm can achieve regret smaller than $\Omega(\sqrt{DSAT})$. (Jaksch et al., 2010) also proposed an optimistic algorithm (UCRL) achieving a regret bound of $O(D\sqrt{\Gamma SAT})$ where $\Gamma \leq S$ is the number of possible next states.[1] While little attention has been devoted to the dependence on D (Bartlett and Tewari, 2009; Talebi and Maillard, 2018; Fruit et al., 2018), several work successfully closed the gap between upper and lower bounds w.r.t. the dependency on the number of states using a PS approach (e.g., Osband et al., 2013; Osband and Roy, 2017; Agrawal and Jia, 2017).[2] Osband and Roy (2017) argued that this improvement is intrinsic in the way PS builds the (high-probability) set of plausible MDPs, which is somehow more accurate than the one built by UCRL.

In this paper, we investigated further the set of plausible MDPs generated by PS algorithms and show that its existing analysis is flawed. In particular, assuming that the reward function is known, it is critical to show that the transition model $p(s'|s,a)$ can be effectively estimated from samples. The regret bounds of PSRL (Osband et al., 2013; Osband and Roy, 2017) and optimistic PSRL (Opt-PSRL) (Agrawal and Jia, 2017) build on a novel concentration inequality for an estimator $\widehat{p}_n$ (either built as a Dirichlet or a Multinomial random variable). We provide both empirical evidence and an asymptotic theoretical argument showing that such concentration inequality may be flawed, thus questioning whether the improvement of $\sqrt{\Gamma}$ in the bound is actually achieveable.

2. Problem Formulation

Provide a minimum level of context about the general proof, just to show where Z_n comes from.

We analyse the concentration properties of the random variable $Z_n \geq 0$ defined as:

$$Z_n := \max_{v \in [0,D]^S} \left\{ (\widehat{p}_n - p)^\mathsf{T} v \right\} \tag{1}$$

where $\widehat{p}_n \in \Delta^S$ is a random vector, $p \in \Delta^S$ is deterministic and $\Delta^S = \{x \in \mathbb{R}^S \ : \ \sum_{i=1}^S x_i = 1 \wedge x_i \geq 0\}$ is the $(S-1)$-dimensional simplex. It is easy to show that the maximum in Eq. 1 is equivalent to computing the (scaled) ℓ_1-norm of the vector $\widehat{p}_n - p$:

$$Z_n = \max_{u \in [-\frac{D}{2}, \frac{D}{2}]} \left\{ (\widehat{p}_n - p)^\mathsf{T} \left(u + \frac{D}{2} e \right) \right\} = \frac{D}{2} \|\widehat{p}_n - p\|_1 \tag{2}$$

1. The original UCRL bound was $O(DS\sqrt{AT})$ but it can be easily refined by considering an empirical Bernstein inequality in place of the Chernoff-Hoeffding (e.g., Fruit et al., 2018).
2. Some of these results only hold for the finite horizon case or Bayesian regret.

where we have used the fact that $\frac{D}{2}(\widehat{p}_n - p)^\mathsf{T} e = 0$. As a consequence, Z_n is a bounded random variable in $[0, D]$. Note that this derivation apply to any random vector $\widehat{p}$, i.e., we can use this argument both for $\widehat{p}_n \sim \textit{Dirichlet}(np)$ (Osband and Roy, 2017) or $\widehat{p}_n \sim \frac{1}{n}\textit{Multinomial}(n, p)$ (Agrawal and Jia, 2017).

Multinomial random variable. We start considering the case of maximum likelihood estimation of the transition probabilities. Formally, let $p(\cdot|s,a)$ be the true (unknown) transition probabilities and denote by $N(s,a,s')$ the number of times a transition $(s,a) \to s'$ was observed. Then, the estimator is $\widehat{p}(s'|s,a) = \frac{N(s,a,s')}{N(s,a)}$ with $N(s,a) = \sum_{s' \in \mathcal{S}} N(s,a,s')$. This is equivalent to say that $\widehat{p}(\cdot|s,a) \sim \frac{1}{n}\textit{Multinomial}(N(s,a), p(\cdot|s,a))$. For sake of clarity, we will focus on a single state-action pair and we will remove the dependence on (s,a).

The literature has analysed the concentration of the ℓ_1-discrepancy of the true distribution and the empirical one in this setting.

Proposition 1 *(Weissman et al., 2003) Let $p \in \Delta^S$ and $\widehat{p} \sim \frac{1}{n}\textit{Multinomial}(n, p)$. Then, for any $S \geq 2$ and $\delta \in [0,1]$:*

$$\mathbb{P}\left(\|\widehat{p} - p\|_1 \geq \sqrt{\frac{2S\ln(2/\delta)}{n}} \right) \leq \mathbb{P}\left(\|\widehat{p} - p\|_1 \geq \sqrt{\frac{2\ln\left((2^S - 2)/\delta\right)}{n}} \right) \leq \delta \tag{3}$$

This concentration inequality is at the core of the proof of UCRL, see (Jaksch et al., 2010, App. C.1). Another inequality is provided in (Devroye, 1983, Lem. 3).

Proposition 2 *(Devroye, 1983) Let $p \in \Delta^S$ and $\widehat{p} \sim \frac{1}{n}\textit{Multinomial}(n, p)$. Then, for any $0 \leq \delta \leq 3\exp(-4S/5)$:*

$$\mathbb{P}\left(\|\widehat{p}_n - p\|_1 \geq 5\sqrt{\frac{\ln(3/\delta)}{n}} \right) \leq \delta \tag{4}$$

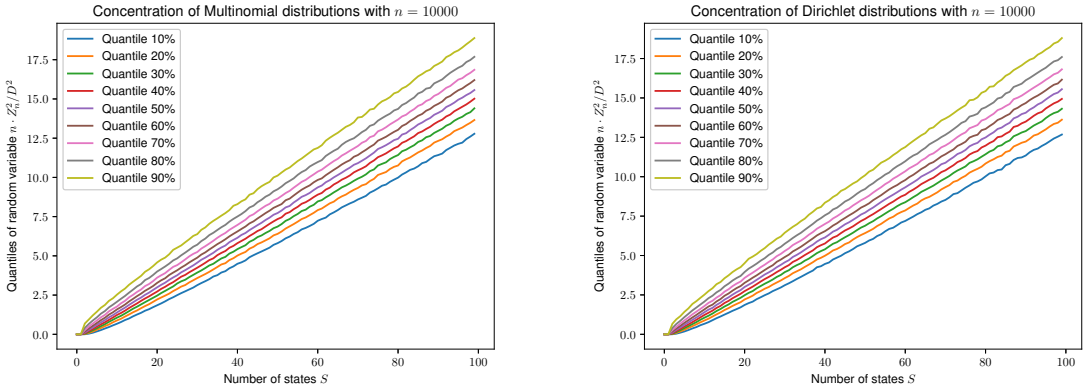
While Prop. 1 shows an explicit dependence on the dimension of the random variable, such dependence is hidden in Prop. 2 by the constraint on δ. Note that for any $0 \leq \delta \leq 3\exp(-4S/5)$, $\sqrt{\frac{\ln(3/\delta)}{n}} > \sqrt{\frac{4S}{5n}}$. This shows that the ℓ_1-deviation always scales proportionally to the dimension of the random variable, i.e., as $\sqrt{S}$.

A better inequality. The natural question is whether is possible to derive a concentration inequality independent from the dimension of p by exploiting the correlation between $\widehat{p}$ and the maximizer vector v^*. This question has been recently addressed in (Agrawal and Jia, 2017, Lem. C.2):

Lemma 3 *(Agrawal and Jia, 2017) Let $p \in \Delta^S$ and $\widehat{p} \sim \frac{1}{n}\textit{Multinomial}(n, p)$. Then, for any $\delta \in [0,1]$:*

$$\mathbb{P}\left(\|\widehat{p}_n - p\|_1 \geq \sqrt{\frac{2\ln(1/\delta)}{n}} \right) \leq \delta$$

Their results resemble the one in Lem. 2 but removes the constraint on δ. As a consequence, the implicit or explicit dependence on the dimension S is removed. In the following, we will show (empirically and theoretically) that Lem. 3 may not be correct.



(a) Multinoulli (b) Dirichlet

Figure 1: Empirical percentiles for the $Z_S^2 = (\lim_{n\to+\infty} \sqrt{n} Z_n)^2$.

Dirichlet random variable. We conclude this section by noticing that a similar concentration inequality has been provided in (Osband and Roy, 2017, Lem. 3) for $\widehat{p} \sim \mathit{Dirichlet}(np)$. Similarly to Lem. 3, the authors showed that $\|\widehat{p} - p\|_1 \leq \sqrt{\frac{2\ln(2/\delta)}{n}}$ with probability $1-\delta$. We will show that even this result may not be correct.

3. Empirical Results

We start providing a simple empirical test that contradicts the result in Lem. 3 and the similar result for Dirichlet random variables. The aim of the test is to compute an empirical estimate of the percentiles of $Z_S = \lim_{n\to+\infty} \sqrt{n} Z_n$.[3]

We set $D = 1$, $n = 10,000$, $p = (\frac{1}{S}, \dots, \frac{1}{S})$ and we let S vary. Then, for each S, we sample $\widehat{p}_j \sim \frac{1}{n}\mathit{Multinomial}(n, p)$ (or $\mathit{Dirichlet}(np)$) and we calculate Z_j as in Eq. 2. The sample $\{\widehat{p}_j\}_{j=1,\dots,n}$ represent the empirical distribution of Z_S. We use this empirical distribution to compute the percentiles for Z_S^2.[4]

We start considering the case of Multinomial random variable. If Lem. 3 holds, we should observe the percentiles to be bounded by a (constant) state-independent value (i.e., $\sqrt{\frac{\alpha\ln(\beta/\delta)}{n}}$). Fig. 1a clearly shows that this is not the case. In particular, the results show a clear *linear* dependence in the parameter S which is in line with the inequality in Lem. 1 and 2. This simple test provides an empirical support to the fact that Lem. 3 may be wrong and, as a consequence, the gain in $\sqrt{S}$ in the regret bound of Opt-PSRL.

A similar conclusion can be derived for the case of Dirichlet random variable as shown in Fig. 1b. This result will invalidate the regret bound of PSRL. Furthermore, it is worth noticing that the empirical distribution of Z_S seems to be equivalent for Multinomial and Dirichlet random variable.

3. In practice, n is chosen to be sufficiently large.
4. We chose to show the squared variable since it is easy to observe a linear dependence w.r.t. to S rather than a square-root one.

4. Theoretical Explanation (the asymptotic case)

In this section, we provide a counter-argument to the Lem. 3 in the asymptotic regime (i.e., $n \to +\infty$). The overall idea is to show that the expected value of Z_n asymptotically grows as $O(\sqrt{S})$ and Z_n itself is well concentrated around its expectation. As a result, we can deduce that all quantiles of Z_n grow as $O(\sqrt{S})$ as well.

Similarly to Sec. 3, we consider the true vector p to be uniform, i.e., $p = (\frac{1}{S}, \dots, \frac{1}{S})$ and $\widehat{p} \sim \frac{1}{n}\mathit{Multinomial}(n, p)$.[5] The following lemma provides a characterization of the variable $Z_S := \lim_{n\to+\infty} \sqrt{n} Z_n$.

Lemma 4 *Consider $S \in \mathbb{N}$, $\mathcal{S} = \{1, \dots, S\}$ and $p = (\frac{1}{S}, \dots, \frac{1}{S})$ be the uniform distribution on $\mathcal{S}$. Let e_S be the vector of ones of dimension S. Define $Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1}N)$ where $N = e_S e_S^\mathsf{T} - I_S$ is the matrix with 0 in all the diagonal entry and 1 elsewhere, and $Y^+ = (\max(Y_i, 0))_{i\in\mathcal{S}}$. Then:*

$$Z_S = \lim_{n\to+\infty} \sqrt{n} Z_n \sim \|Y^+\|_1 D \sqrt{\frac{S-1}{S^2}}.$$

Furthermore,

$$\mathbb{E}[Z_S] = \sqrt{\frac{S-1}{S^2}} \cdot \mathbb{E}\left[\sum_{i=1}^{S} Y_i^+\right] = \sqrt{S-1} \cdot \mathbb{E}[Y_1^+] = \sqrt{\frac{S-1}{2\pi}}.$$

While the previous lemma may already suggest that Z_S should grow as $O(\sqrt{S})$ as its expectation, it is still possible that a large part of the distribution is concentrated around a value independent from S, with limited probability assigned to, e.g., values growing as $O(S)$, which could justify the $O(\sqrt{S})$ growth of the expectation. Thus, in order to conclude the analysis, we need to show that Z_S is concentrated "enough" around its expectation.

Since the random variables Y_i are correlated, it is complicated to directly analyze the deviation of Z_S from its mean. Thus we first apply an orthogonal transformation on Y to obtain independent r.v. (recall that jointly normally distributed variables are independent if uncorrelated).

Lemma 5 *Consider the same settings of Lem. 4 and recall that $Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1}N)$. There exists an orthogonal transformation $U \in O_S(\mathbb{R})$, s.t.*

$$W = \sqrt{\frac{S-1}{S}} UY \sim \mathcal{N}\left(0, \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}\right).$$

By exploiting the transformation U we can write that $Z_S \sim g(W) := \frac{1}{\sqrt{S}} e_S^\mathsf{T} \left(U^\mathsf{T} W\right)^+$. Since W_i are i.i.d. standard Gaussian random variables and g is 1-Lipschitz, we can finally characterize the mean and the deviations of Z_S and derive the following anticoncentration inequality for Z_S.

5. The analysis holds also in the case $\widehat{p} \sim \mathit{Direchlet}(np)$.

Theorem 6 *Let* $p \in \Delta^S = (\frac{1}{S}, \dots, \frac{1}{S})$ *and* $\widehat{p}_n \sim \frac{1}{n}$*Multinomial*$(n, p)$*. Define* $Z_n = \max_{v \in [0,D]} \left\{ (\widehat{p}_n - p)^\mathsf{T} v \right\}$ *and* $Z_S = \lim_{n \to +\infty} \sqrt{n} Z_n$*. Then, for any* $\delta \in (0,1)$*:*

$$\mathbb{P}\Big[Z_S \geq \sqrt{\frac{2(S-1)}{\pi}} - \sqrt{2\log(2/\delta)} \Big] \geq 1 - \delta.$$

This result shows that every quantile of Z_S is dependent on the dimension of the random variable, i.e., $\sqrt{S}$. Similarly to Lem. 2, it is possible to lower bound the quantile by a dimension-free quantity at the price of having an exponential dependence on S in δ. We want to stress again that this is not removing the dependence on the dimension but simply hiding it. Finally, this result is coherent with the empirical test we have reported in Sec. 3.

Remarks. Note that the application of the presented analysis to the case of $\widehat{p} \sim$ *Dirichlet*(np) leads to the same result, thus invalidating the results in (Osband and Roy, 2017).

Finally, we have devoted App. B to present the possible mistakes in the proof of Lem. 3.

5. Conclusion

We have presented arguments supporting a possible mistake in the analysis of PS algorithms. This paper shows that there may be a statical limitation for achieving a $\sqrt{S}$-dependence in the regret bound using current proof techniques. It still might be possible to improve the dependence on the number of states by using the specific structure of RL problems, e.g., the fact that v is related to $\widehat{p}$ by value iteration. In particular, Azar et al. (2017) have achieved a $\sqrt{S}$ dependence by leveraging on the finite-horizon structure and the value iteration process.

Finally, it is still not clear if it will be possible to achieve a simple $\sqrt{S}$ dependence in the regret bound in place of $\sqrt{\Gamma S}$. In particular, in the construction of the lower bound, Jaksch et al. (2010) treated Γ as a constant value since in any MDP they considered $\Gamma = 2$. An analysis and refinement of the lower bound can shed light on this open question.

Appendix A. Proof for the asymptotic scenario

In this section we report the proofs of lemmas and theorem stated in Sec. 4.

A.1 Proof of Lem. 4

Let $Y_{n,i} = \frac{1}{\sqrt{n}\sqrt{\frac{S-1}{S^2}}} \sum_{j=1}^{n} (X_i^j - \frac{1}{S})$ and $Y_n = (Y_{n,i})_{i \in \mathcal{S}}$. Then:

$$\begin{aligned}\sqrt{n} Z_n &= \sqrt{n} \max_{v \in [0,D]^S} (\widehat{p} - p)^\mathsf{T} v = \sqrt{n} \max_{v \in [0,D]^S} \sum_{i=1}^{S} \frac{v_i}{n} \sum_{j=1}^{n} (X_i^j - \frac{1}{S}) \\ &= \max_{v \in [0,D]^S} \sum_{i=1}^{S} Y_{n,i} v_i \sqrt{\frac{S-1}{S^2}} = D \sqrt{\frac{S-1}{S^2}} \cdot e^\mathsf{T} Y_n^+,\end{aligned}$$

where we used the fact that the v maximizing Z_n takes the largest value D for all positive components $Y_{n,i}$ and is equal to 0 otherwise. We recall that the covariance of the normalized multinoulli variable $Y_{n,i}$ with probabilities $p_i = 1/S$ is $I_S - \frac{1}{S-1} N$. As a result, a direct application of the central limit theorem gives $Y_n \xrightarrow{\mathcal{D}} Y \sim \mathcal{N}(0, I_S - \frac{1}{S-1} N)$. Then we can apply the functional CLT and obtain $Z_S = \lim_{n \to \infty} \sqrt{n} Z_n = \lim_{n \to \infty} e_S^\mathsf{T} Y_n^+ \sqrt{\frac{S-1}{S^2}} \overset{\mathcal{D}}{\sim} \sqrt{\frac{S-1}{S^2}} \cdot e^\mathsf{T} Y^+$, where Y^+ is a random vector obtained by truncating from below at 0 the multi-variate Gaussian vector Y. Since the marginal distribution of each random variable Y_i is $\mathcal{N}(0,1)$, i.e., are identically distributed (see definition in Lem. 4), Y_i^+ has a distribution composed by a Dirac distribution in 0 and a half normal distribution, and its expected value is $\mathbb{E}[Y_i^+] = 1/\sqrt{2\pi}$, while leads to the final statement on the expectation.

A.2 Proof of Lem. 5

Denote $\lambda(A)$ the set of eigenvalues of square matrix A. Let $B \in \mathbb{R}^{S \times S}$ such that $B = \begin{bmatrix} 0_{S,S-1} & e_S \end{bmatrix}$, where $0_{S,S-1} \in \mathbb{R}^{S \times (S-1)}$ is a matrix full of zeros. Then, we can write the eigenvalues of the covariance matrix of Y as

$$\begin{aligned}\lambda(I_S - \frac{1}{S-1} N) &= \lambda(\frac{S}{S-1} I_S - \frac{1}{S-1} e_S e_S^\mathsf{T}) = \lambda\left(\frac{S}{S-1} I_S - \frac{1}{S-1} B B^\mathsf{T}\right) \\ &= \frac{S}{S-1} \lambda\left(I_S - \frac{1}{S-1} B^\mathsf{T} B\right) = \frac{S}{S-1} \lambda\left(I_S - \begin{bmatrix} 0_{S-1} & 0 \\ 0 & 1 \end{bmatrix}\right),\end{aligned}$$

where we use the fact that $\lambda(I - A^\mathsf{T} A) = \lambda(I - A A^\mathsf{T})$. As a result, the covariance of Y has one eigenvalue at 0 and eigenvalues equal to $\frac{S}{S-1}$ with multiplicity $S-1$. As a result, we can diagonalize it with an orthogonal matrix $U \in O_S(\mathbb{R})$ (obtained using the normalized eigenvectors) and obtain

$$U(I_S - \frac{1}{S-1} N) U^T = \begin{bmatrix} \frac{S}{S-1} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Define $W = \sqrt{\frac{S-1}{S}}UY$, then:

$$Cov(W,W) = \frac{S-1}{S}Cov(UY,UY) = \frac{S-1}{S}UCov(Y,Y)U^T$$
$$= \frac{S-1}{S}U(I_S - \frac{1}{S-1}N)U^T = \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}.$$

Thus $W \sim \mathcal{N}\left(0, \begin{bmatrix} I_{S-1} & 0 \\ 0 & 0 \end{bmatrix}\right)$.

A.3 Proof of Thm. 6

By exploiting Lem. 4 and Lem. 5 we can write:

$$Z_S \sim e_S^\mathsf{T} Y^+ \cdot \sqrt{\frac{S-1}{S^2}} = e_S^\mathsf{T}\left(\sqrt{\frac{S}{S-1}}U^\mathsf{T}W\right)^+ \cdot \sqrt{\frac{S-1}{S^2}} = e_S^\mathsf{T}\left(U^\mathsf{T}W\right)^+ \cdot \frac{1}{\sqrt{S}}$$

Let $g(\cdot) = e_S^\mathsf{T}\left(U^T\cdot\right)^+ \frac{1}{\sqrt{S}}$. Then g is 1-Lipschitz:

$$|g(x) - g(y)| \leq Lip(e_s^\mathsf{T}\cdot)Lip(U^\mathsf{T}\cdot)Lip((\cdot)^+)\frac{1}{\sqrt{S}}\|x-y\|_2 = \sqrt{S}\cdot 1\cdot 1\cdot\frac{1}{\sqrt{S}}\|x-y\|_2$$

where $Lip(f)$ denotes the Lipschitz constant of a function f and we exploit the fact that U is an orthonormal matrix.

We can now study the concentration of the variable Z_S. Given that W is a vector of i.i.d. standard Gaussian variables[6] and g is 1-Lipschitz, we can use (Wainwright, 2017, Thm. 2.4) to prove that for all $t > 0$:

$$\mathbb{P}(Z_S \geq \mathbb{E}[Z_S] - t) \geq 1 - \mathbb{P}(|Z_S - \mathbb{E}[Z_S]| \geq t) \geq 1 - 2e^{-\frac{t^2}{2}}.$$

Substituting the value of $\mathbb{E}[Z_S]$ and inverting the bound gives the desired statement.

Appendix B. Comments on the proof of Lem. 3

This section aims to give an idea of the possible mistakes in the proof of Lem. 3 (Agrawal and Jia, 2017). We give for granted the knowledge of the proof structure since we point out very specific problems

The first step of the proof is to define $\{Y_v\}$ *indipendent* binomial random variables distributed as $Y_v \sim \frac{1}{n}Binomial(n, \frac{1}{D}p^\mathsf{T}v)$. Then DY_v is stochastically optimistic compared to $\widehat{p}^\mathsf{T}v$. However, we believe that it is not possible to claim that $\mathbb{E}[DY_v - \widehat{p}^\mathsf{T}v|\widehat{p}^\mathsf{T}v] = 0$, as the authors do. In particular, according to Cond. 3 in Lem. 3 (Osband et al., 2014), we can only claim that there is a random variable $W \sim DY_v$ such that $\mathbb{E}[W - \widehat{p}^\mathsf{T}v|\widehat{p}^\mathsf{T}v] = 0$. The proof will require a change in the random variable has shown by the following example.

Example 1 *Let $X \sim \mathcal{N}(0,1), Y = -X$, then it is clear that $X \sim Y$, thus $X \geq_{so} Y$, but $\mathbb{E}[X - Y|Y] = -2Y \neq 0$.*

6. Note that we can drop the last component of W since it is deterministically zero.

One quick fix for this part of the proof is to take Y_v as constructed in Lem. E.8 and Cor. E.9 in (Agrawal and Jia, 2017). Nevertheless this choice breaks the independence assumption on $\{Y_v\}$ as proved in the following example.

Example 2 *If we take $Y_v = \hat{q}_v, Y_w = \hat{q}_w$ as in Lemma E.8, then if we take $n = 1$, we will have*

$$\mathbb{P}(\hat{q}_v = 1) = \mathbb{P}(\sum_{j=1}^{S} X^j Y_v^j = 1) = \sum_{j=1}^{S} \mathbb{P}(X^j = 1)\mathbb{P}(Y_v^j = 1) = \sum_{j=1}^{S} p_j v_j / D.$$

Similarly, $\mathbb{P}(\hat{q}_w = 1) = \sum_{j=1}^{S} p_j w_j / D.$

$$\begin{aligned}\mathbb{P}(\hat{q}_v = 1, \hat{q}_w = 1) &= \mathbb{P}(\sum_{j=1}^{S} X^j Y_v^j = 1, \sum_{j=1}^{S} X^j Y_w^j = 1) \\ &= \sum_{j=1}^{S} \mathbb{P}(X^j = 1, Y_v^j = 1, Y_w^j = 1) = \sum_{j=1}^{S} p_j \frac{v_j}{D} \frac{w_j}{D} \\ &\neq \mathbb{P}(\hat{q}_v = 1)\mathbb{P}(\hat{q}_w = 1) = (\sum_{j=1}^{S} p_j v_j / D)(\sum_{k=1}^{S} p_k w_k / D).\end{aligned}$$

Without independence is still possible to claim the stochastic dominance but the rest of the proof will not go through. In particular, without independence, the distribution of Y is more intricate than stated. The authors defined $Y(w) = \int_v \mathbb{1}(Z(w) \in \mathcal{E}_v) Y_v(w)$. Note that $Z(w)$ is correlated to $\widehat{p}$ thus to X_i^j thus to Y_v. Then, we are not able to claim as in the proof that $\mathbb{1}(\tilde{v} = v, Y_v^j) \sim \mathbb{1}(\tilde{v} = v)(Ber(\mu_v) - \mu_v)$.

There may exist different ways to tackle the proof but, as there are arguments supporting the invalidity of the result, we believe that there may a mistake in the proof related to a not correct use of Lem. E.8 (Agrawal and Jia, 2017) and Cond.3 in Lem. 3 (Osband et al., 2014). A similar argument applies to the case of Dirichlet distribution (e.g., Osband and Roy, 2017; Agrawal and Jia, 2017).

References

Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In *UAI*, pages 1–11. AUAI Press, 2015.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *NIPS*, pages 1184–1194, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 263–272, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Peter L. Bartlett and Ambuj Tewari. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *UAI*, pages 35–42. AUAI Press, 2009.

Luc Devroye. The equivalence of weak, strong and complete convergence in ℓ_1 for kernel density estimates. *The Annals of Statistics*, 11(3):896–904, 09 1983. doi: 10.1214/aos/1176346255.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. Regret minimization in mdps with options without prior knowledge. In *NIPS*, pages 3169–3179, 2017.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *CoRR*, abs/1802.04020, 2018.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *ICML*, 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 *NIPS*, pages 3003–3011, 2013.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *CoRR*, abs/1402.0635, 2014.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *NIPS*, pages 1333–1342, 2017.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in mdps. In *ALT*, volume 83 of *Proceedings of Machine Learning Research*, pages 770–805. PMLR, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. 2017.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Labs, 2003.