Optimistic Thompson Sampling-based Algorithms for Episodic Reinforcement Learning

Bingshan Hu[1,2] **Tianyue H. Zhang**[3] **Nidhi Hegde**[1,4] **Mark Schmidt**[3,4]

[1]Department of Computing Science, University of Alberta, Canada
[2]Alberta Machine Intelligence Institute (Amii), Canada
[3]Department of Computer Science, University of British Columbia, Canada
[4]Canada CIFAR AI Chair, Alberta Machine Intelligence Institute (Amii), Canada

Abstract

We propose two Thompson Sampling-like, model-based learning algorithms for episodic Markov decision processes (MDPs) with a finite time horizon. Our proposed algorithms are inspired by Optimistic Thompson Sampling (O-TS), empirically studied in Chapelle and Li [2011], May et al. [2012] for stochastic multi-armed bandits. The key idea for the original O-TS is to clip the posterior distribution in an optimistic way to ensure that the sampled models are always better than the empirical models. Both of our proposed algorithms are easy to implement and only need one posterior sample to construct an episode-dependent model. Our first algorithm, Optimistic Thompson Sampling for MDPs (O-TS-MDP), achieves a $\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$ regret bound, where S is the size of the state space, A is the size of the action space, H is the number of time-steps per episode and T is the number of episodes. Our second algorithm, Optimistic Thompson Sampling plus for MDPs (O-TS-MDP$^+$), achieves the (near)-optimal $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ regret bound by taking a more aggressive clipping strategy. Since O-TS was only empirically studied previously, we derive regret bounds of O-TS for stochastic bandits. In addition, we propose, O-TS-Bandit$^+$, a randomized version of UCB1 [Auer et al., 2002], for stochastic bandits. Both O-TS and O-TS-Bandit$^+$ achieve the optimal $O\left(\frac{A\ln(T)}{\Delta}\right)$ problem-dependent regret bound, where Δ denotes the sub-optimality gap.

1 INTRODUCTION

Reinforcement learning (RL) algorithms have been widely implemented in real-world applications such as autonomous driving, image processing, natural language processing, financial modeling, gaming, etc. Typically, an RL task can be formulated as a Markov decision process (MDP) with a state space, an action space, and a state transition function. In each time-step, the learning agent visits a state, plays an action, and transitions to the next state. In this paper, we consider the learning problem of episodic, non-stationary MDPs with S states, A actions, and a finite time horizon H. In each round $t = 1, 2, \ldots, H$ belonging to episode $k = 1, 2, \ldots, T$, the learning agent visits a state and plays an action according to an action-sampling strategy. Then, the learning agent receives a random reward drawn from a fixed but unknown reward distribution and transitions to the next state sampled from a fixed but unknown transition probability distribution associated with the played action. The goal of the learning agent is to take actions wisely to maximize the cumulative reward over T episodes. The learning agent faces an exploitation-vs-exploration dilemma. In a single round, the learning agent can only choose an action that empirically performs the best so far to maximize the cumulative reward (exploitation) or choose an action that has not been played too often to learn the parameters of the associated unknown distributions (exploration).

Upper Confidence Bound (UCB)-based algorithms, inspired by the philosophy of optimism in the face of uncertainty (OFU), can achieve a balance between exploitation and exploration. The high-level idea behind this class of algorithms is to construct upper confidence bounds by adding an extra term to the empirical estimates. The additive term encourages the learning agent to play actions that have not been played too often. Many existing episodic MDP learning algorithms [Azar et al., 2017, Dann et al., 2017, Zanette and Brunskill, 2019, Dann et al., 2019, Zhang et al., 2020, Tiapkin et al., 2022b] are UCB-based. Notably, two model-based algorithms, UCBVI [Azar et al., 2017] and Bayes-UCBVI [Tiapkin et al., 2022b], enjoy the (near)-optimal $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ regret bound.[1] UCB-based algorithms usu-

[1] The $\widetilde{O}(\cdot)$ notation only hides poly $\log(ASHT)$ factors.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

ally do not randomize the obtained data, and the exploration is driven by the additive terms.

Another class of algorithms perturb the obtained data in a certain way to encourage the learning agent to visit states and actions that have been less explored. The extra randomness can be achieved by injecting noise into the data. As shown in Osband et al. [2019], learning algorithms with random noise efficiently drive exploration. Many prominent state-of-the-art algorithms [Russo, 2019, Pacchiano et al., 2021, Xiong et al., 2022, Agrawal et al., 2021] for episodic MDPs are developed by adding calibrated Gaussian noise to the empirical estimates of the rewards. Although the transition probability distribution is unknown, the aforementioned learning algorithms do not add any noise to the empirical estimates of the transition probability distributions. Among them, a model-based algorithm, NARL-UCBVI [Pacchiano et al., 2021], and a model-free algorithm, C-RLSVI [Agrawal et al., 2021], achieve the same $\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$ regret bound. Very recently, SSR-Bernstein [Xiong et al., 2022], also a model-free learning algorithm, tightens the regret bound to $\widetilde{O}\left(\sqrt{ASH^3T}\right)$.

Different from adding noise directly to the data, extra randomness can also be achieved by drawing random samples from well-constructed data-dependent distributions. Interestingly, if the data-dependent distribution is designed to be the posterior distribution that models an unknown parameter, it coincides with the description of Thompson Sampling [Thompson, 1933, Chapelle and Li, 2011], one of the oldest randomized algorithms. For example, Gaussian distributions can be used to model the means of the reward distributions and Dirichlet distributions can be used to model the transition probability distributions in MDPs.

Thompson Sampling was originally invented for stochastic bandits, which can be viewed as a simple MDP with only one state, A actions, and one round per episode. Since there are no transitions between states, the only unknown parameters in a stochastic bandit problem are the means of the reward distributions. Conceptually, Thompson Sampling plays an action according to the posterior probability distribution of the optimal action. Empirically, it is not required to compute the exact posterior probability distribution of the optimal action. Instead, Thompson Sampling can draw a random sample from the posterior distribution associated with each action and then play the action with the highest posterior sample. The practical performance of Thompson Sampling has been studied extensively by Chapelle and Li [2011]. The theoretical performance of Thompson Sampling relies on the choice of the prior distributions. Thompson Sampling with Beta priors is asymptotically optimal [Agrawal and Goyal, 2017, Kaufmann et al., 2012], while Thompson Sampling with Gaussian priors is problem-dependent optimal [Agrawal and Goyal, 2017].

In reality, it is not necessary to restrict the data-dependent distribution to the exact posterior distribution. If the data-dependent distribution is designed to be the posterior distribution with some parameters modified, we term this type of algorithm as *Thompson Sampling-like algorithms with posterior distribution reshaping*. Optimistic Thompson Sampling (O-TS), a Thompson Sampling-like algorithm with posterior distribution reshaping, was first introduced and empirically evaluated by Chapelle and Li [2011], May et al. [2012].[2] The key idea of O-TS is to boost the random posterior sample to the mean of the posterior distribution (Gaussian distribution) if the random sample is smaller than the mean. In other words, O-TS reshapes the posterior distribution from a Gaussian distribution to a one-sided Gaussian distribution with the left side being clipped. There are other Thompson Sampling-like algorithms with reshaped posterior distributions. For stochastic bandits, Jin et al. [2021] devise MOTS, the first Thompson Sampling-like algorithm achieving minimax optimality. MOTS reshapes the posterior distribution by both clipping the upper tail and boosting the variance of the posterior distribution.

Different from stochastic bandits where the learning agent only needs to learn the means of the reward distributions, in episodic MDPs, the transition probability distributions are also unknown. Two Thompson Sampling-like, model-based algorithms, SOS-OPS-RL [Agrawal and Jia, 2020] and OPSRL [Tiapkin et al., 2022a], use Dirichlet distributions to model the transition probability distributions. To drive exploration, they also boost the variance of the Dirichlet distributions. SOS-OPS-RL achieves a $\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$ regret bound while SPSRL achieves the (near)-optimal $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ regret bound.

Now, we list our key contributions in this paper.

(1) We propose O-TS-MDP, a computationally efficient and theoretically elegant model-based learning algorithm with randomized value functions, for episodic MDPs. O-TS-MDP only draws one random sample and enjoys a $\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$ regret bound. There are two key ingredients in O-TS-MDP. The first one is the usage of the boosted variance of the posterior distribution (a Gaussian distribution) to drive exploration. The second one is the usage of O-TS to clip the left side of the posterior distribution to simplify the theoretical analysis. Although the regret bound of O-TS-MDP is not as tight as OPSRL [Tiapkin et al., 2022a] and SSR-Bernstein [Xiong et al., 2022], OPSRL needs to draw $\widetilde{O}(1)$ random samples while SSR-Bernstein is a model-free algorithm. For our regret analysis of O-TS-MDP, we can avoid upper bounding the absolute value of the estimation error, thus simplifying the theoretical analysis as compared to the analysis of RLSVI-based algorithms [Russo, 2019, Agrawal et al., 2021, Xiong et al., 2022].

[2] Originally, this learning algorithm was called Optimistic Bayesian Sampling (OBS) [May et al., 2012].

(2) We propose O-TS-MDP$^+$, a model-based, OFU-inspired optimistic algorithm with randomized value functions.[3] O-TS-MDP$^+$ achieves the (near)-optimal $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ regret bound. The key idea in O-TS-MDP$^+$ is a more aggressive clipping strategy of the posterior distribution. O-TS-MDP$^+$ boosts the value of the random sample to the upper confidence bound if the random sample is smaller than the upper confidence bound. The aggressive clipping contributes to reducing the variance of the reshaped posterior distribution as compared to O-TS-MDP. Consequently, the regret bound is tightened to be (near)-optimal. O-TS-MDP$^+$ can be viewed as a randomized version of UCB-VI [Azar et al., 2017].

(3) Although Chapelle and Li [2011], May et al. [2012] have demonstrated the empirical performance of O-TS for stochastic bandits, there is no theoretical analysis for it. We derive regret bounds for O-TS for bandits. In addition, we propose O-TS-Bandit$^+$, an OFU-inspired learning algorithm, for stochastic bandits. Both O-TS and O-TS-Bandit$^+$ achieve the (order)-optimal $O\left(\frac{A\ln(T)}{\Delta}\right)$ problem-dependent regret bound, where Δ is the sub-optimality gap.

2 LEARNING PROBLEM

We consider an episodic non-stationary MDP problem which can be specified by $M = \{\mathcal{S}, \mathcal{A}, H, \boldsymbol{P}, \boldsymbol{\mu}, p_0\}$, where $\mathcal{S}$ is a finite state-space with size S, $\mathcal{A}$ is a finite action-space with size A, H is the finite number of rounds in each episode and p_0 is the deterministic initial state distribution. Let $\boldsymbol{P}$ and $\boldsymbol{\mu}$ denote the transition function and reward function, respectively. The learning agent interacts with the environment in an episodic way with the following learning protocol. In each round $t \in [H]$ belonging to episode k, the learning agent observes a state s_t^k and plays an action a_t^k. Then, the learning agent receives a random reward $X^k_{s_t^k, a_t^k, t} \in [0, 1]$ that is drawn from a fixed reward distribution with mean $\mu_{s_t^k, a_t^k, t} = \boldsymbol{\mu}(s_t^k, a_t^k, t)$ and transitions to the next state s_{t+1}^k that is sampled from a fixed transition probability distribution $P_{s_t^k, a_t^k, t} = \boldsymbol{P}(s_t^k, a_t^k, t)$. The initial state s_1^k is sampled from p_0. The goal of the learning agent is to accumulate as much reward as possible over a finite number of T episodes, i.e., HT rounds in total.

A deterministic policy $\pi = (\pi(\cdot, 1), \pi(\cdot, 2), \ldots, \pi(\cdot, H))$ is a sequence of functions, where each $\pi(\cdot, t) : \mathcal{S} \to \mathcal{A}$ takes a state as input and outputs an action that will be played if the learning agent visits that state. Let Π collect all such policies. The value function $V_t^\pi(s)$ for state s of policy π in round t is defined as $V_t^\pi(s) := \mu_{s,\pi(s,t),t} + P^{\intercal}_{s,\pi(s,t),t} V^\pi_{t+1}$. If we knew $\boldsymbol{P}$ and $\boldsymbol{\mu}$, we could use backwards induction to compute the optimal policy π_*. Define $V^{\pi_*}_{H+1} = \vec{0}$. Then, for each round $t = H, H-1, \ldots, 1$, for each state $s \in \mathcal{S}$, we compute

$$\begin{aligned} \pi_*(s,t) &= \arg\max_{a \in \mathcal{A}} \left\{ \mu_{s,a,t} + P^{\intercal}_{s,a,t} V^{\pi_*}_{t+1} \right\} \quad , \\ V_t^{\pi_*}(s) &= \mu_{s,\pi_*(s,t),t} + P^{\intercal}_{s,\pi_*(s,t),t} V^{\pi_*}_{t+1} \quad . \end{aligned}$$

The expected regret $\mathcal{R}(T)$ over T episodes is defined as

$$\mathcal{R}(T) = \sum_{k=1}^{T} \mathbb{E}\left[\left(V_1^{\pi_*}(s_1^k) - V_1^{\pi_k}(s_1^k)\right)\right] \quad , \qquad (1)$$

where π_k is random and the initial state $s_1^k \sim p_0$. If there exists (s, t) such that $\pi_k(s,t) \neq \pi_*(s,t)$, then regret may occur. Define $\mathcal{F}_k = \left\{ s_t^q, a_t^q, X^q_{s_t^q, a_t^q, t}, t \in [H], q \in [k] \right\}$ as the history trajectory by the end of episode k following policies $\pi_1, \ldots, \pi_q, \ldots, \pi_k$ with the initial state in each episode independently drawn from p_0. Define $\mathcal{F}_0 = \{\}$.

3 O-TS-MDP AND O-TS-MDP$^+$

Algorithm 1 O-TS-MDP

```
Input: MDP instance $M$, number of episodes $T$
Initialization:
Set $\widehat{O}_{s,a,t} \leftarrow 0$, $\widehat{P}_{s,a,t} \leftarrow \vec{0}$, $\widehat{\mu}_{s,a,t} \leftarrow 0, \forall (s,a,t)$
for episode $k = 1, 2, \ldots, T$ do
    Set $\widetilde{V'}^{\pi_k}_{H+1} = \vec{0}$
    for $t = H, H-1, \ldots, 1$ do
        for $s \in \mathcal{S}$ do
            for $a \in \mathcal{A}$ do
                Draw $\widetilde{\mu}_{s,a,t} \sim \mathcal{N}\left(\widehat{\mu}_{s,a,t}, \left(\sqrt{SH}\sigma^k_{s,a,t}\right)^2\right)$
                Set $\widetilde{\mu}'_{s,a,t} \leftarrow \max\left\{\widetilde{\mu}_{s,a,t}, \widehat{\mu}_{s,a,t}\right\}$
                Set $\widetilde{Q}_{s,a,t} \leftarrow \widetilde{\mu}'_{s,a,t} + \widehat{P}^{\intercal}_{s,a,t} \widetilde{V'}^{\pi_k}_{t+1}$
            end for
            Set $\pi_k(s,t) \leftarrow \arg\max_{a \in \mathcal{A}} \widetilde{Q}_{s,a,t}$
            Set $\widetilde{V'}^{\pi_k}_{t}(s) \leftarrow \widetilde{Q}_{s,\pi_k(s,t),t}$
        end for
    end for
    Sample $s_1^k \sim p_0$, run $\pi_k$, and update $\widehat{\mu}_{s_t^k, \pi_k(s_t^k,t), t}$,
    $\widehat{O}_{s_t^k, \pi_k(s_t^k,t), t}$, and $\widehat{P}_{s_t^k, \pi_k(s_t^k,t), t}$ for all $t \in [H]$.
end for
```

We first present some notations used by our proposed algorithms O-TS-MDP and O-TS-MDP$^+$. Let $\widehat{O}^k_{s,a,t} = \sum_{q=1}^{k} \mathbf{1}\left\{(s_t^q, a_t^q) = (s,a)\right\}$ denote the number of times that (s,a) has been visited in round t by the end of episode k. Let $\widehat{\mu}^k_{s,a,t} = \frac{1}{\widehat{O}^k_{s,a,t}} \sum_{q=1}^{k} \mathbf{1}\left\{(s_t^q, a_t^q) = (s,a)\right\} X^q_{s,a,t}$ denote the empirical mean of (s,a,t) by the end of episode k. Let $\widehat{P}^k_{s,a,t}(s') = \frac{1}{\widehat{O}^k_{s,a,t}} \sum_{q=1}^{k} \mathbf{1}\left\{(s_t^q, a_t^q) = (s,a), s^q_{t+1} = s'\right\}$ denote the empirical transition probability distribution. Let $\sigma^k_{s,a,t} := 5\sqrt{H^2 \log^2\left(\frac{H}{\delta}\right) / \widehat{O}^{k-1}_{s,a,t}}$, where $\delta = \frac{1}{ASH^2T^2}$.

[3] We say a learning algorithm is OFU-inspired and optimistic if the Optimism Decomposition [Pacchiano et al., 2021] can be used to decompose the regret.

For the case when (s, a, t) has not been visited yet by the end of episode $k-1$, i.e., $\widehat{O}^{k-1}_{s,a,t} = 0$, our algorithms set $\sigma^k_{s,a,t}$ to a large constant.

3.1 O-TS-MDP

O-TS-MDP is presented in Algorithm 1. The key ingredients in O-TS-MDP are the boosted variance of the posterior distribution (a Gaussian distribution) to drive exploration and the usage of O-TS [May et al., 2012, Chapelle and Li, 2011] to clip a Gaussian distribution to a one-sided Gaussian distribution with the left side being truncated. The reshaping of the posterior distribution plays a crucial role in simplifying the algorithm and the theoretical analysis.

The same as other model-based algorithms, in episode k, O-TS-MDP constructs an episode-dependent model $\tilde{M}'_k$ to simulate the true model. To construct $\tilde{M}'_k$ with randomized value functions, O-TS-MDP draws a random sample $\widetilde{\mu}^k_{s,a,t} \sim \mathcal{N}\left(\widehat{\mu}^{k-1}_{s,a,t}, SH\left(\sigma^k_{s,a,t}\right)^2\right)$ for each (s, a, t). If $\widetilde{\mu}^k_{s,a,t} < \widehat{\mu}^{k-1}_{s,a,t}$, O-TS-MDP boosts it to $\widehat{\mu}^{k-1}_{s,a,t}$. Let $\widetilde{\mu'}^k_{s,a,t} := \max\left\{\widehat{\mu}^{k-1}_{s,a,t}, \widetilde{\mu}^k_{s,a,t}\right\}$. Note that $\widetilde{\mu'}^k_{s,a,t}$ can be viewed as a random variable drawn from distribution $\mathcal{N'}^k_{s,a,t}$ with the probability density function (PDF) $f'(x) =$

$$\begin{cases} 0, & x < \widehat{\mu}^{k-1}_{s,a,t}, \\ \phi\left(x; \widehat{\mu}^{k-1}_{s,a,t}, SH\left(\sigma^k_{s,a,t}\right)^2\right) + \frac{\delta\left(x-\widehat{\mu}^{k-1}_{s,a,t}\right)}{2}, & x \geq \widehat{\mu}^{k-1}_{s,a,t}, \end{cases}$$

where $\phi(x; \mu, \sigma^2)$ denotes the PDF of $\mathcal{N}\left(\mu, \sigma^2\right)$ and $\delta(\cdot)$ denotes the Dirac delta function.[4] With $\widetilde{\mu'}^k_{s,a,t}$ for all (s, a, t) in hand, we construct $\tilde{M}'_k = \left\{\mathcal{S}, \mathcal{A}, H, \widehat{P}^{k-1}, \widetilde{\mu'}^k, p_0\right\}$, where $\widehat{P}^{k-1} = \left\{\widehat{P}^{k-1}_{s,a,t}\right\}$ collects all the empirical transition probability distributions by the end of episode $k-1$ and $\widetilde{\mu'}^k = \left\{\widetilde{\mu'}^k_{s,a,t}\right\}$ collects all the random samples after the boosting. After constructing $\tilde{M}'_k$, O-TS-MDP uses backwards induction to find the optimal policy π_k for $\tilde{M}'_k$ (shown in Line 4 to Line 12 in Algorithm 1). Let $\widetilde{V'}^{\pi}_t$ denote the value functions of a fixed policy π for $\tilde{M}'_k$ in round t.

Now, we present a regret bound for Algorithm 1.

Theorem 1. *The regret of Algorithm 1 is* $\widetilde{O}\left(\sqrt{AS^2H^4T}\right)$.

O-TS-MDP is a *computationally efficient and space efficient* algorithm which only needs one random sample for each (s, a, t) to construct the episode-dependent model. Per episode, the time complexity is $O(AS^2H)$ and the space complexity is $O(AS^2H)$. In contrast, OPSRL [Tiapkin et al., 2022a] and SOS-OPS-RL [Agrawal and Jia, 2020] need multiple posterior samples to construct a model. The improvement of O-TS-MDP comes from the usage of $\widehat{P}^{k-1}$ to construct the model. Instead, OPSRL and SOS-OPS-RL use Dirichlet random variables to construct the model. Although O-TS-MDP, OPSRL, and SOS-OPS-RL are all model-based algorithms with randomized value functions, O-TS-MDP is not an optimistic learning algorithm while OPSRL and SOS-OPS-RL are both optimistic algorithms. In other words, in O-TS-MDP, the value functions $\widetilde{V'}^{\pi_*}_1(s)$ are not guaranteed to be greater than $V^{\pi_*}_1(s)$ with high probability. As shown in the regret analysis, O-TS-MDP only achieves weak optimism. That is, each $\widetilde{V'}^{\pi_*}_1(s)$ are only guaranteed to be greater than $V^{\pi_*}_1(s)$ with a small constant probability. However, the strong optimism guarantee in OPSRL and SOS-OPS-RL is at the cost of drawing multiple posterior samples. We believe that the regret bound of O-TS-MDP can also be tightened to $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ if drawing $\widetilde{O}(1)$ random samples.

The key idea behind SSR-Bernstein [Xiong et al., 2022] to achieve the optimal $\widetilde{O}\left(\sqrt{ASH^3T}\right)$ regret bound is to limit the amount of randomness within the learning algorithm. More specifically, in SSR-Bernstein, in each episode, all tuples (s, a, t) use the same random seed, which is a Gaussian random variable. In other words, for the entire learning algorithm across T episodes, the number of independent Gaussian random variables in SSR-Bernstein is exactly T. In contrast, in O-TS-MDP, each tuple (s, a, t) has its own random seed meaning that the total amount of independent Gaussian random variables is $O(ASHT)$. As compared to O-TS-MDP, SSR-Bernstein does not fully randomize its obtained data.

Although O-TS-MDP and the RLSVI-based algorithms [Russo, 2019, Agrawal et al., 2021, Xiong et al., 2022] share in common an exploration mechanism, the introduce of O-TS to clip the left side of the posterior distribution *simplifies the theoretical analysis in two ways*. First, upper bounding the *absolute value of the estimation error* is not needed in the theoretical analysis of O-TS-MDP (Proof of Lemma 2 in the appendix presents more details). All the aforementioned RLSVI-based algorithms upper bound the absolute value of the estimation error, which is more complicated than upper bounding the one-sided error, as stated in Xiong et al. [2022]. Second, as compared to C-RLSVI of Agrawal et al. [2021] and SSR of Xiong et al. [2022], O-TS-MDP does not clip the randomized value functions to values in $[0, H]$. Consequently, the analysis of O-TS-MDP can reuse the value difference lemma (Lemma 16) directly.

Recall that $\sigma^k_{s,a,t} = 5\sqrt{H^2 \log^2(T/\delta)/\widehat{O}^{k-1}_{s,a,t}}$. To sketch the regret analysis, we first construct the empirical MDP $\hat{M}_k = \left\{\mathcal{S}, \mathcal{A}, H, \widehat{P}^{k-1}, \widehat{\mu}^{k-1}, p_0\right\}$, where $\widehat{\mu}^{k-1} =$

[4] Note that from Algorithm 1 we can see that to implement O-TS-MDP, it is not required to compute $f'(x)$. We simply draw a random sample from a Gaussian distribution and then compare the sample value with the mean of the Gaussian distribution.

$\left\{\widehat{\mu}_{s,a,t}^{k-1}\right\}$ collects all the empirical means. Let $\widehat{V}_t^{\pi}$ denote the value functions of a fixed policy π for $\hat{M}_k$. To decompose the regret, we define two high-probability events:

$$\begin{aligned}
\mathcal{E}^k \quad &= \quad \left\{\left|\left(\widehat{P}_{s,a,t}^{k-1} - P_{s,a,t}\right)^{\intercal} V_{t+1}^{\pi_k}\right| \leq \sigma_{s,a,t}^k,\right. \\
&\qquad \left.\left|\left(\widehat{\mu}_{s,a,t}^{k-1} - \mu_{s,a,t}\right)\right| \leq \sigma_{s,a,t}^k, \forall (s,a,t)\right\}, \\
\mathcal{E}_{\pi_*}^k \quad &= \quad \left\{\left|\left(P_{s,a,t} - \widehat{P}_{s,a,t}^{k-1}\right)^{\intercal} V_{t+1}^{\pi_*}\right| \leq \sigma_{s,a,t}^k\right. \\
&\qquad \left.\left|\widehat{\mu}_{s,a,t}^{k-1} - \mu_{s,a,t}\right| \leq \sigma_{s,a,t}^k, \forall (s,a,t)\right\}.
\end{aligned} \tag{2}$$

We prepare three lemmas to prove Theorem 1. Recall $V_1^{\pi_*}(s)$ is the value function of the optimal policy π_* for the true MDP M and $\widetilde{V'}_1^{\pi_k}(s)$ is the value function of the optimal policy π_k for the episode-dependent MDP $\tilde{M}_k'$. Our technical Lemma 2 upper bounds the expected performance gap between the optimal policy π_* over M and the optimal policy π_k over $\tilde{M}_k'$.

Lemma 2. *(Optimism). In episode k, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left(V_1^{\pi_*}(s_1^k) - \widetilde{V'}_1^{\pi_k}(s_1^k)\right) \mathbf{1}\left\{\mathcal{E}_{\pi_*}^k\right\}\right] \\
\leq\ &\sqrt{SH} \sum_{t=1}^{H} \mathbb{E}\left[O\left(\sigma_{s_t^k, \pi_k(s_t^k, t), t}^k\right)\right].
\end{aligned} \tag{3}$$

Recall that $\widehat{V}_1^{\pi_k}(s)$ is the value function of policy π_k over the empirical MDP $\hat{M}_k$. Lemma 3 upper bounds the expected performance gap for a single policy π_k over MDPs $\tilde{M}_k'$ and $\hat{M}_k$ and Lemma 4 upper bounds the expected performance gap for policy π_k over MDPs $\hat{M}_k$ and M. Note that MDPs $\tilde{M}_k'$ and $\hat{M}_k$ are constructed based on the same $\widehat{P}^{k-1}$ which is determined by $\mathcal{F}_{k-1}$.

Lemma 3. *(Posterior deviation). In episode k, we have*

$$\mathbb{E}\left[\widetilde{V'}_1^{\pi_k}(s_1^k) - \widehat{V}_1^{\pi_k}(s_1^k)\right] \leq \sqrt{SH} \sum_{t=1}^{H} \mathbb{E}\left[\sigma_{s_t^k, \pi_k(s_t^k, t), t}^k\right]. \tag{4}$$

Lemma 4. *(Empirical deviation). In episode k, we have*

$$\begin{aligned}
&\mathbb{E}\left[\left(\widehat{V}_1^{\pi_k}(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \mathbf{1}\left\{\mathcal{E}^k\right\}\right] \\
&\leq 2 \sum_{t=1}^{H} \mathbb{E}\left[\sigma_{s_t, \pi_k(s_t, t), t}^k\right].
\end{aligned} \tag{5}$$

Proof of Theorem 1. We have

$$\begin{aligned}
&\sum_{k=1}^{T} \mathbb{E}\left[\left(V_1^{\pi_*}(s_1^k) - V_1^{\pi_k}(s_1^k)\right)\right] \\
\leq\ &\sum_{k=1}^{T} \mathbb{E}\left[\underbrace{\left(V_1^{\pi_*}(s_1^k) - \widetilde{V'}_1^{\pi_k}(s_1^k)\right) \mathbf{1}\left\{\mathcal{E}_{\pi_*}^k\right\}}_{\text{Lemma 2}}\right] \\
+\ &\sum_{k=1}^{T} \mathbb{E}\left[\underbrace{\left(\widetilde{V'}_1^{\pi_k}(s_1^k) - \widehat{V}_1^{\pi_k}(s_1^k)\right)}_{\text{Lemma 3}}\right] \\
+\ &\sum_{k=1}^{T} \mathbb{E}\left[\underbrace{\left(\widehat{V}_1^{\pi_k}(s_1^k) - V_1^{\pi_k}(s_1^k)\right) \mathbf{1}\left\{\mathcal{E}^k\right\}}_{\text{Lemma 4}}\right] \\
+\ &H \sum_{k=1}^{T} \underbrace{\mathbb{P}\left\{\overline{\mathcal{E}^k}\right\} + \mathbb{P}\left\{\overline{\mathcal{E}_{\pi_*}^k}\right\}}_{\text{Lemma 18}} \\
\leq\ &\sqrt{SH} \cdot \mathbb{E}\left[\sum_{k=1}^{T} \sum_{t=1}^{H} O\left(\sigma_{s_t^k, \pi_k(s_t^k, t), t}^k\right)\right] + O(1) \\
\leq\ &\widetilde{O}\left(\sqrt{AS^2H^4T}\right),
\end{aligned} \tag{6}$$

where the last inequality uses $\sum_{k=1}^{T} \sum_{t=1}^{H} O\left(\sigma_{s_t^k, \pi_k(s_t^k, t), t}^k\right) \leq \widetilde{O}\left(\sqrt{ASH^3T}\right)$, a well-known result in the MDP literature [Russo, 2019, Agrawal et al., 2021]. □

3.2 O-TS-MDP$^+$

Different from O-TS-MDP where only weak optimism is guaranteed, O-TS-MDP$^+$ is an OFU-inspired optimistic algorithm with randomized value functions. O-TS-MDP$^+$ takes a more aggressive clipping strategy to achieve strong optimism. O-TS-MDP$^+$ boosts the random sample to the upper confidence bound if it is smaller than the upper confidence bound. This aggressive clipping strategy contributes to reducing the variance of the posterior distribution as compared to O-TS-MDP, which, consequently, leads to tightening the regret bound to $\widetilde{O}\left(\sqrt{ASH^3T}\right)$.

Similar to Algorithm 1, in each episode, O-TS-MDP$^+$ also constructs a model $\tilde{M}_k' = \left\{\mathcal{S}, \mathcal{A}, H, \widehat{P}^{k-1}, \widetilde{\mu'}^k, p_0\right\}$. Now, we present how to construct $\widetilde{\mu'}^k$. Let $\overline{\mu}_{s,a,t}^k := \widehat{\mu}_{s,a,t}^{k-1} + 2\sigma_{s,a,t}^k$ be the upper confidence bound for (s,a,t) which is determined by $\mathcal{F}_{k-1}$. At the beginning of episode k, for each (s,a,t), O-TS-MDP$^+$ draws a random sample $\widetilde{\mu}_{s,a,t}^k \sim \mathcal{N}\left(\widehat{\mu}_{s,a,t}^{k-1}, \left(\sigma_{s,a,t}^k\right)^2\right)$. Then, O-TS-MDP$^+$ boosts it to $\overline{\mu}_{s,a,t}^k$ if $\widetilde{\mu}_{s,a,t}^k < \overline{\mu}_{s,a,t}^k$. Let $\widetilde{\mu'}_{s,a,t}^k = \max\left\{\widetilde{\mu}_{s,a,t}^k, \overline{\mu}_{s,a,t}^k\right\}$ denote the sample value after the boosting. The PDF for the distribution of $\widetilde{\mu'}_{s,a,t}^k$ can be defined as $f'(x) = 0$ if $x < \overline{\mu}_{s,a,t}^k$. Otherwise, $f'(x) =$

$\phi\left(x; \widehat{\mu}^{k-1}_{s,a,t}, (\sigma_{s,a,t})^2\right) + \Phi\left(\overline{\mu}^k_{s,a,t}; \widehat{\mu}^{k-1}_{s,a,t}, (\sigma_{s,a,t})^2\right)\delta(x - \overline{\mu}^k_{s,a,t})$, where $\Phi(x;\mu,\sigma^2)$ denotes the cumulative distribution function (CDF) of $\mathcal{N}\left(\mu,\sigma^2\right)$. Let $\widetilde{\mu'}^k = \left\{\widetilde{\mu'}^k_{s,a,t}\right\}$ collect all the samples after the boosting. After constructing $\tilde{M}'_k$, O-TS-MDP$^+$ computes the optimal policy π_k for $\tilde{M}'_k$ by using backwards induction. Algorithm 2 presents the pseudo-code of O-TS-MDP$^+$. The differences between O-TS-MDP and O-TS-MDP$^+$ are highlighted in Algorithm 1 and Algorithm 2, respectively.

Algorithm 2 O-TS-MDP$^+$

```
Input: MDP instance $M$, number of episodes $T$
Initialization:
Set $\widehat{O}_{s,a,t} \leftarrow 0$, $\widehat{P}_{s,a,t} \leftarrow \vec{0}$, $\widehat{\mu}_{s,a,t} \leftarrow 0, \forall (s,a,t)$
for episode $k = 1, 2, \ldots, T$ do
    Set $\widetilde{V'}^{\pi_k}_{H+1} = \vec{0}$
    for $t = H, H-1, \ldots, 1$ do
        for $s \in \mathcal{S}$ do
            for $a \in \mathcal{A}$ do
                Draw $\widetilde{\mu}_{s,a,t} \sim \mathcal{N}\left(\widehat{\mu}_{s,a,t}, \left(\sigma^k_{s,a,t}\right)^2\right)$
                Set $\overline{\mu}_{s,a,t} \leftarrow \widehat{\mu}_{s,a,t} + 2\sigma^k_{s,a,t}$
                Set $\widetilde{\mu}'_{s,a,t} \leftarrow \max\left\{\widetilde{\mu}_{s,a,t}, \overline{\mu}_{s,a,t}\right\}$
                Set $\widetilde{Q}_{s,a,t} \leftarrow \widetilde{\mu}'_{s,a,t} + \widehat{P}^{\mathsf{T}}_{s,a,t}\widetilde{V'}^{\pi_k}_{t+1}$
            end for
            Set $\pi_k(s,t) \leftarrow \arg\max_{a\in\mathcal{A}} \widetilde{Q}_{s,a,t}$
            Set $\widetilde{V'}^{\pi_k}_t(s) \leftarrow \widetilde{Q}_{s,\pi_k(s,t),t}$
        end for
    end for
    Sample $s^k_1 \sim p_0$, run $\pi_k$, and update $\widehat{\mu}_{s^k_t,\pi_k(s^k_t,t),t}$,
    $\widehat{O}_{s^k_t,\pi_k(s^k_t,t),t}$ and $\widehat{P}_{s^k_t,\pi_k(s^k_t,t),t}$ for all $t \in [H]$.
end for
```

Now, we present a regret bound for Algorithm 2.

Theorem 5. *The regret of Algorithm 2 is* $\widetilde{O}\left(\sqrt{ASH^3T}\right)$.

O-TS-MDP$^+$ achieves the same (near)-optimal regret bound as OPSRL of Tiapkin et al. [2022a] and SSR-Bernstein of Xiong et al. [2022]. O-TS-MDP$^+$ can be viewed as a randomized version of UCB-VI [Azar et al., 2017]. It is important to note that the O-TS-MDP$^+$ learning algorithm itself does not need a Bernstein-type bonus. However, the analysis of Theorem 5 relies on a concentration bound that is derived based on Bernstein's inequality (see Lemma 17 for more details).

Now, we sketch the regret analysis. Recall that $\overline{\mu}^k_{s,a,t} = \widehat{\mu}^{k-1}_{s,a,t} + 2\sigma^k_{s,a,t}$. We first construct a new MDP $\bar{M}_k = \left\{\mathcal{S}, \mathcal{A}, H, \widehat{P}^{k-1}, \overline{\mu}^k, p_0\right\}$, where $\overline{\mu}^k = \left\{\overline{\mu}^k_{s,a,t}\right\}$ collects all the upper confidence bounds. Let $\overline{V}^\pi_t$ be the value functions of a fixed policy π for $\bar{M}_k$. Note that conditioned on history $\mathcal{F}_{k-1}$, although the constructed $\bar{M}_k$ is deterministic, $\overline{V}^{\pi_k}_1(s)$ is still random as π_k is random. Recall that π_k is the optimal policy for $\tilde{M}'_k$. We still use $\mathcal{E}^k$ and $\mathcal{E}^k_{\pi_*}$, the events that have been defined in (2), to decompose the regret. We prepare three lemmas for Theorem 5.

Lemma 6. *(Optimism). In episode k, we have*

$$\mathbb{E}\left[\left(V^{\pi_*}_1(s^k_1) - \widetilde{V'}^{\pi_k}_1(s^k_1)\right)\mathbf{1}\left\{\mathcal{E}^k_{\pi_*}\right\}\right] \leq 0. \tag{7}$$

Lemma 7. *(Posterior deviation). In episode k, we have*

$$\mathbb{E}\left[\left(\widetilde{V'}^{\pi_k}_1(s^k_1) - \overline{V}^{\pi_k}_1(s^k_1)\right)\right] \leq \sum_{t=1}^{H}\mathbb{E}\left[\sigma^k_{s^k_t,\pi_k(s^k_t,t),t}\right]. \tag{8}$$

Lemma 8. *(UCB-like). In episode k, we have*

$$\begin{aligned}&\mathbb{E}\left[\left(\overline{V}^{\pi_k}_1(s^k_1) - V^{\pi_k}_1(s^k_1)\right)\mathbf{1}\left\{\mathcal{E}^k\right\}\right]\\ \leq\ & 2\sum_{t=1}^{H}\mathbb{E}\left[\sigma^k_{s^k_t,\pi_k(s^k_t,t),t}\right].\end{aligned} \tag{9}$$

Proof of Theorem 5. We have

$$\begin{aligned}&\sum_{k=1}^{T}\mathbb{E}\left[\left(V^{\pi_*}_1(s^k_1) - V^{\pi_k}_1(s^k_1)\right)\right]\\ \leq\ & \sum_{k=1}^{T}\mathbb{E}\left[\underbrace{\left(V^{\pi_*}_1(s^k_1) - \widetilde{V'}^{\pi_k}_1(s^k_1)\right)\mathbf{1}\left\{\mathcal{E}^k_{\pi_*}\right\}}_{\text{Optimism, Lemma 6}}\right]\\ +\ & \sum_{k=1}^{T}\mathbb{E}\left[\underbrace{\left(\widetilde{V'}^{\pi_k}_1(s^k_1) - \overline{V}^{\pi_k}_1(s^k_1)\right)}_{\text{Posterior deviation, Lemma 7}}\right]\\ +\ & \sum_{k=1}^{T}\mathbb{E}\left[\underbrace{\left(\overline{V}^{\pi_k}_1(s^k_1) - V^{\pi_k}_1(s^k_1)\right)\mathbf{1}\left\{\mathcal{E}^k\right\}}_{\text{UCB-like, Lemma 8}}\right]\\ +\ & H\sum_{k=1}^{T}\underbrace{\mathbb{P}\left\{\overline{\mathcal{E}^k}\right\} + \mathbb{P}\left\{\overline{\mathcal{E}^k_{\pi_*}}\right\}}_{\text{Lemma 18}}\\ \leq\ & \mathbb{E}\left[\sum_{k=1}^{T}\sum_{t=1}^{H}O\left(\sigma^k_{s^k_t,\pi_k(s^k_t,t),t}\right)\right] + O(1)\\ \leq\ & \widetilde{O}\left(\sqrt{ASH^3T}\right).\end{aligned} \tag{10}$$

□

4 O-TS AND O-TS-BANDIT$^+$

Since a stochastic bandit problem can be viewed as a special MDP with $S = 1, H = 1$ and Chapelle and Li [2011] have already demonstrated the empirical performance of O-TS for stochastic bandits, to fill a gap in the stochastic bandit literature we present regret bounds of O-TS for stochastic bandits. In addition, we propose O-TS-Bandit$^+$, an OFU-inspired, optimistic learning algorithm, for stochastic bandits. Note that O-TS-Bandit$^+$ can be viewed as a randomized version of UCB1 [Auer et al., 2002].

Now, we present the learning problem of stochastic bandits with bounded rewards formally. In a stochastic bandit problem, we have an arm set $\mathcal{A}$ with size A. At the beginning of each round t, the environment generates a reward vector $X(t) = (X_1(t), X_2(t), \ldots, X_A(t))$ with each $X_j(t) \in [0, 1]$ i.i.d. over time from a fixed but unknown probability distribution with mean μ_j. Simultaneously, the learning agent pulls an arm $J_t \in \mathcal{A}$. At the end of round t, the learning agent observes and obtains $X_{J_t}(t)$, the reward associated with the pulled arm. The goal of the learning agent is to pull arms sequentially to accumulate as much reward as possible over a finite number of T rounds. Without loss of generality, we assume that the first arm is the unique optimal arm. In other words, we assume $\mu_1 > \mu_j$ for all $j \neq 1$. Let $\Delta_j := \mu_1 - \mu_j$ denote the mean reward gap.

We use regret to measure the performance of the learning agent's decisions. Similar to (1), the regret is defined as

$$\mathcal{R}(T) \quad = \quad T \cdot \mu_1 - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{J_t}\right] \quad , \qquad (11)$$

where the expectation is taken over J_t. Different from episodic MDPs, where only the worst-case regret bounds are analyzed, for stochastic bandits we are interested in both problem-dependent regret bounds and problem-independent regret bounds. The difference between problem-dependent regret bounds and problem-independent regret bounds is the former one depends on the mean reward parameters $\mu_1, \mu_2, \ldots, \mu_A$ while the latter one provides a regret bound for all the possible choices of mean reward parameters.

We introduce additional notation specific to stochastic bandits. Let $O_j(t-1)$ denote the number of pulls of arm j by the end of round $t-1$ and $\widehat{\mu}_{j,O_j(t-1)}$ denote the empirical mean of arm j by the end of round $t-1$.

4.1 O-TS

The learning algorithm of O-TS is presented in Algorithm 3. Similar to Algorithm 1, at the beginning of each round t, for each arm $j \in \mathcal{A}$, a random sample $\widetilde{\mu}_j(t)$ is drawn from $\mathcal{N}\left(\widehat{\mu}_{j,O_j(t-1)}, 1/O_j(t-1)\right)$. If $\widetilde{\mu}_j(t)$ is smaller than $\widehat{\mu}_{j,O_j(t-1)}$, it will be boosted to $\widehat{\mu}_{j,O_j(t-1)}$. Let $\widetilde{\mu}'_j(t) = \max\left\{\widetilde{\mu}_j(t), \widehat{\mu}_{j,O_j(t-1)}\right\}$. With all $\widetilde{\mu}'_j(t)$ in hand, the learning agent pulls arm $J_t = \arg\max_{j \in \mathcal{A}} \widetilde{\mu}'_j(t)$.

Now, we present regret bounds for Algorithm 3.

Theorem 9. *The problem-dependent regret bound of Algorithm 3 is* $\sum_{j \in \mathcal{A}: \Delta_j > 0} O\left(\frac{\ln(T)}{\Delta_j}\right)$.

Theorem 10. *The problem-independent regret bound of Algorithm 3 is* $O\left(\sqrt{AT\ln(A)}\right)$.

Algorithm 3 O-TS (Optimistic Thompson Sampling [Chapelle and Li, 2011])

```
Input: an arm set $\mathcal{A}$
Pull each arm once to initialize $O_j, \widehat{\mu}_{j,O_j}$
for round $t = A+1, A+2, \ldots$ do
    for $a \in \mathcal{A}$ do
        Draw $\widetilde{\mu}_j(t) \sim \mathcal{N}\left(\widehat{\mu}_{j,O_j}, 1/O_j\right)$
        Set $\widetilde{\mu}'_j(t) \leftarrow \max\left\{\widetilde{\mu}_j(t), \widehat{\mu}_{j,O_j}\right\}$
    end for
    Pull arm $J_t \leftarrow \arg\max_{j \in \mathcal{A}} \widetilde{\mu}'_j(t)$
    Update $O_{J_t}$ and $\widehat{\mu}_{J_t,O_{J_t}}$ .
end for
```

O-TS achieves the same problem-dependent and problem-independent regret bounds as Thompson Sampling with Gaussian priors (TS-Gaussian) (Algorithm 2 in Agrawal and Goyal [2017]). The key difference between O-TS and TS-Gaussian is TS-Gaussian uses normal distributions while O-TS uses one-sided Gaussian distributions with the left side being clipped. The problem-independent regret bound of O-TS is minimax optimal up to a $\sqrt{\ln(A)}$ factor. Note that O-TS and TS-Gaussian are not optimistic learning algorithms.

4.2 O-TS-BANDIT$^+$

An OFU-inspired optimistic learning algorithm, O-TS-Bandit$^+$, is presented in Algorithm 4. Similar to Algorithm 2, O-TS-Bandit$^+$ does the clipping aggressively to boost optimism and can be viewed as a randomized version of UCB1 [Auer et al., 2002]. Let $\overline{\mu}_j(t) = \widehat{\mu}_{j,O_j(t-1)} + \sqrt{1.5\ln(t)/O_j(t-1)}$ be the upper confidence bound. O-TS-Bandit$^+$ boosts $\widetilde{\mu}_j(t)$ to $\overline{\mu}_j(t)$ if it is smaller than $\overline{\mu}_j(t)$. Let $\widetilde{\mu}'_j(t) = \max\left\{\widetilde{\mu}_j(t), \overline{\mu}_j(t)\right\}$ denote the value after the boosting. Then, O-TS-Bandit$^+$ pulls arm $J_t = \arg\max_{j \in \mathcal{A}} \widetilde{\mu}'_j(t)$. The differences between O-TS and O-TS-Bandit$^+$ are highlighted in Algorithm 3 and Algorithm 4, respectively.

Algorithm 4 O-TS-Bandit$^+$

```
Input: an arm set $\mathcal{A}$
Pull each arm once to initialize $O_j, \widehat{\mu}_{j,O_j}$
for round $t = A+1, A+2, \ldots$ do
    for $j \in \mathcal{A}$ do
        Draw $\widetilde{\mu}_j(t) \sim \mathcal{N}\left(\widehat{\mu}_{j,O_j}, 1/O_j\right)$
        Set $\overline{\mu}_j \leftarrow \widehat{\mu}_{j,O_j} + \sqrt{1.5\ln(t)/O_j}$
        Set $\widetilde{\mu}'_j(t) \leftarrow \max\left\{\widetilde{\mu}_j(t), \overline{\mu}_j(t)\right\}$
    end for
    Pull arm $J_t \leftarrow \arg\max_{j \in \mathcal{A}} \widetilde{\mu}'_j(t)$
    Update $O_{J_t}$ and $\widehat{\mu}_{J_t,O_{J_t}}$ .
end for
```

Now, we present regret bounds for Algorithm 4.

Theorem 11. *The problem-dependent regret bound of Algorithm 4 is* $\sum_{j\in\mathcal{A}:\Delta_j>0} O\left(\frac{\ln(T)}{\Delta_j}\right)$.

Theorem 12. *The problem-independent regret bound of Algorithm 4 is* $O\left(\sqrt{AT\ln(T)}\right)$.

O-TS and O-TS-Bandit$^+$ have the same problem-dependent regret bound. For the problem-independent regret bound, O-TS-Bandit$^+$ is worse than O-TS. Note that O-TS-Bandit$^+$ is an optimistic algorithm. O-TS-Bandit$^+$ and MOTS [Jin et al., 2021] share in common that they both clip the Gaussian distributions based on the upper confidence bounds. The key difference lies in that MOTS clips the upper tail of Gaussian distributions to control the overestimation of the sub-optimal arms while O-TS-Bandit$^+$ keeps the upper tail of the Gaussian distributions to preserve optimism.

5 EXPERIMENTAL RESULTS

In this section, we evaluate the empirical performance of our proposed algorithms O-TS-MDP and O-TS-MDP$^+$ for MDPs with $S = [5, 20, 50]$, $A = 3$ and $H = 10$. For a fair performance comparison, our experimental set-up is fully adopted from Dann et al. [2017], where the empirical performance for several UCB-based algorithms was studied. For a specific (s, a, t), the random reward $X_{s,a,t}$ in each episode is drawn from a Bernoulli distribution with parameter $\mu_{s,a,t}$. To ensure the sparsity of the random rewards, we set $\mu_{s,a,t} = 0$ with probability 0.85 and with probability 0.15, the value of $\mu_{s,a,t}$ is drawn from a uniform distribution. The sparsity design is to control the occurrence that sub-optimal policies can obtain rewards by chance. We compare O-TS-MDP, O-TS-MDP$^+$, SSR-Bernstein [Xiong et al., 2022], and TS-MDP, a Thompson Sampling-based learning algorithm without clipping the posterior distributions, i.e., constructing the episode-dependent model as $\tilde{M} = \left\{\mathcal{S}, \mathcal{A}, H, \widehat{P}^{k-1}, \tilde{\mu}^k, p_0\right\}$. We set $T = 10^7$ and compare the cumulative average rewards of each episode.

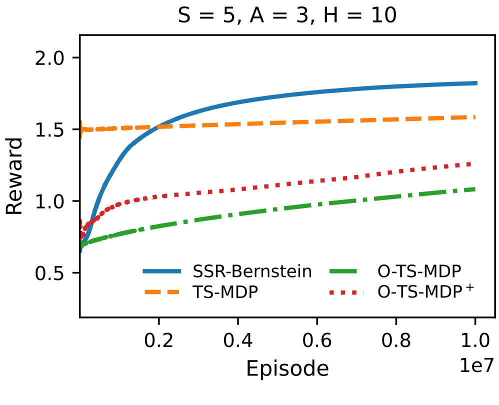

Figure 1: Empirical performance for 5 states

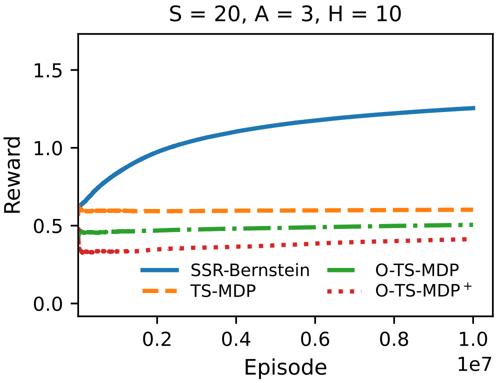

Figure 2: Empirical performance for 20 states

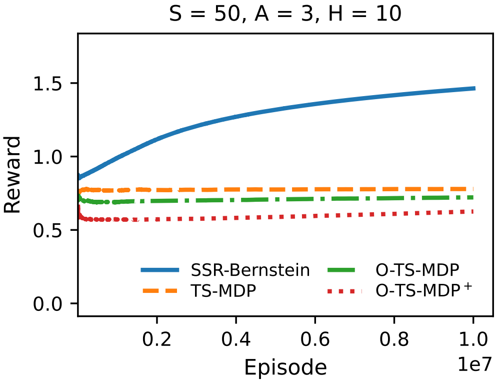

Figure 3: Empirical performance for 50 states

As shown in Figure 1, the rewards for all algorithms steadily increase as the learning agent gains a better estimation of the parameters of the true MDP over time. When the number of states is small ($S = 5$), O-TS-MDP$^+$ performs slightly better than O-TS-MDP. TS-MDP demonstrates a similar trend as O-TS-MDP since they are both Thompson Sampling-based algorithms. Despite the lack of theoretical analysis, TS-MDP does achieve better empirical performance. The gap between O-TS-MDP and TS-MDP comes from the fact that clipping the left side of the posterior distributions increases the chance to visit a sub-optimal (s, a, t), just as implied in the design of MOTS [Jin et al., 2021]. It is not surprising that SSR-Bernstein outperforms the remaining algorithms as it is theoretically optimal. Figure 2 and 3 show similar trends for Thomson-sampling-based methods in larger state space. SSR-Bernstein still performs the best.

It is important to note that SSR-Bernstein uses a single random seed for all (s, a, t) in each episode. In contrast, in our proposed algorithms, each (s, a, t) has its own randomness within an episode. In other words, our proposed algorithms inject more randomness than SSR-Bernstein. Additionally, SSR-Bernstein needs to construct confidence intervals to tune the magnitude of the variance, whereas our algorithms are simpler and easier to implement and enjoy good regret bounds. We also implement UCB-VI [Azar et al., 2017] and more experimental results can be found in Appendix H.

6 CONCLUSION AND FUTURE WORK

In this work, we have presented two Optimistic Thompson Sampling-based learning algorithms, O-TS-MDP and O-TS-MDP$^+$, for episodic MDPs. The key feature that distinguishes our proposed learning algorithms from the existing RLSVI-based algorithms [Russo, 2019, Xiong et al., 2022, Agrawal et al., 2021] is the introduction of O-TS to avoid upper bounding the absolute value of the estimation error, thus simplifying the regret analysis. This work leaves two interesting open questions. Just as pointed out in Abeille and Lazaric [2017], Pacchiano et al. [2021], Agrawal et al. [2021], removing the extra $\sqrt{SH}$ factor is challenging if the learning algorithms are Thompson Sampling-based. The first open question is whether the ideas in SSR of Xiong et al. [2022], i.e., controlling the amount of randomness within the learning algorithm, can be used to tighten the regret bound of O-TS-MDP to $\widetilde{O}\left(\sqrt{ASH^3T}\right)$. Our thought is that by reducing the amount of posterior random samples, a better regret bound for O-TS-MDP may be possible. The analysis of O-TS-MDP and RLSVI-based algorithms all rely on the property that the sum of multiple independent normal random variables is still normally distributed, and normal distributions have nice anti-concentration bounds. Although Tiapkin et al. [2022a] have proved a sharp anti-concentration bound for Dirichlet distributions, the distribution of the sum of multiple independent Dirichlet random variables is still less understood. The lack of understanding of Dirichlet distributions results in the need for multiple posterior samples in OPSRL of Tiapkin et al. [2022a]. The second interesting open question is whether we can reshape the Dirichlet posterior distribution in an optimistic way to improve the number of Dirichlet random variables in OPSRL to one.

ACKNOWLEDGEMENTS

We thank Alan Milligan for helping with the experiments and providing comments on an early version of the manuscript. This work was partially supported by the Alberta Machine Intelligence Institute (Amii), the Canada CIFAR AI Program and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant RGPIN-2022-03669.

References

Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Priyank Agrawal, Jinglin Chen, and Nan Jiang. Improved worst-case regret bounds for randomized least-squares value iteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6566–6573, 2021.

Shipra Agrawal and Navin Goyal. Near-optimal regret bounds for Thompson Sampling. http://www.columbia.edu/~sa3305/papers/j3-corrected.pdf, 2017.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: Worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2020.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine learning*, 47:235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Olivier Chapelle and Lihong Li. An empirical evaluation of Thompson Sampling. *Advances in neural information processing systems*, 24, 2011.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516. PMLR, 2019.

Tianyuan Jin, Pan Xu, Jieming Shi, Xiaokui Xiao, and Quanquan Gu. MOTS: Minimax optimal Thompson Sampling. In *International Conference on Machine Learning*, pages 5074–5083. PMLR, 2021.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, pages 199–213. Springer, 2012.

Benedict C May, Nathan Korda, Anthony Lee, and David S Leslie. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research*, 13:2069–2106, 2012.

Ian Osband, Benjamin Van Roy, Daniel J Russo, Zheng Wen, et al. Deep exploration via randomized value functions. *J. Mach. Learn. Res.*, 20(124):1–62, 2019.

Aldo Pacchiano, Philip Ball, Jack Parker-Holder, Krzysztof Choromanski, and Stephen Roberts. Towards tractable optimism in model-based reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1413–1423. PMLR, 2021.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Mark Rowland, Michal Valko, and Pierre MENARD. Optimistic posterior sampling for reinforcement learning with few samples and tight guarantees. In *Advances in Neural Information Processing Systems*, 2022a.

Daniil Tiapkin, Denis Belomestny, Éric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, and Pierre Ménard. From Dirichlet to Rubin: Optimistic exploration in RL without bonuses. In *International Conference on Machine Learning*, pages 21380–21431. PMLR, 2022b.

Zhihan Xiong, Ruoqi Shen, Qiwen Cui, Maryam Fazel, and Simon Shaolei Du. Near-optimal randomized exploration for tabular Markov decision processes. In *Advances in Neural Information Processing Systems*, 2022.

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*, pages 7304–7312. PMLR, 2019.

Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. *Advances in Neural Information Processing Systems*, 33:15198–15207, 2020.