Worst-Case Regret Bound for Perturbation Based Exploration in Reinforcement Learning

Ziping Xu
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
zipingxu@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48109
tewaria@umich.edu

Abstract

Methods that add random perturbations to estimates before making decisions were first introduced in online learning where they called FTPL (Follow the Perturbed Leader) methods. These methods solve decision making problems under uncertainty by repeatedly solving optimization problems on perturbed input instances. In this paper, we apply this type of perturbation to the design of reinforcement learning algorithm. We provide the first worst-case regret bound for a perturbation algorithm under the unichain, infinite horizon, tabular setting. We believe this study provides a new direction for the analysis of non-optimistic exploration. We also hope that our work further strengthens the connections between optimization and reinforcement learning.

1 Introduction

Designing exploration strategies that are both computationally and statistically efficient is an important problem in reinforcement learning (RL). There have been rich theoretical studies balancing exploration and exploitation in Markov decision processes (MDPs) with finite state and action space (i.e., the tabular setting) (Jaksch et al., 2010; Osband and Van Roy, 2017; Dann and Brunskill, 2015; Kearns and Singh, 2002).

Most previous theoretical analyses utilize optimism in face of uncertainty (OFU), which maintains a confidence set containing true MDP and optimizes over all MDPs in the set. OFU is difficult to use in practical RL settings, in which ϵ-greedy and Boltzmann exploration are commonly used. Many papers attempt to analyze randomized exploration method without optimism. Osband et al. (2013); Osband and Van Roy (2017) study posterior sampling algorithm that samples an MDP from its posterior distribution and executes its optimal policy. They show many advantages over OFU algorithms in terms of computational efficiency. However, they only give a Bayesian regret bound that measures expected regret with respect to a prior distribution over MDPs. As far as we know, Russo (2019) gives the first worst-case regret bound for randomized exploration. They use randomized least squares value iteration (RLSVI) under episodic MDP setting and the crucial proof step is to show a significant probability of being optimistic.

Perturbation algorithm adds random perturbation to estimates before solving optimization problem. As the most straightforward exploration method, there have been many studies on perturbation methods in the bandit literature (Kalai and Vempala, 2005; Kujala and Elomaa, 2005; Van Erven et al., 2014), while it has not been studied well in the RL community. The method has low computational complexity and is closer to commonly used methods in practical settings, i.e. ϵ-greedy exploration. Kim and Tewari (2019) gives a general perturbation framework and also shows the importance of significant probability of optimism. The probability of being optimistic in bandit problems

Preprint. Under review.

corresponds to the tail distribution of the noise that can be easily analyzed, while in RL problems, perturbed transition probability has a complex effect on the average reward.

Our main contribution is to provide the first worst-case regret bound for infinite horizon unichain MDPs by directly perturbing the estimate of transition probability toward the same but random direction for all state-action pairs. Our algorithm, analogous to that of Russo (2019), only perturbs once and the most crucial step is also to show that under our perturbation there is a significant probability of being optimistic. We extend the perturbation framework from bandit to RL problems and we believe this is crucial to analyze more general perturbation methods, for example, posterior sampling.

2 Background on RL

RL problem. We consider the classical tabular Markov decision process (MDP) setting. Let $M = \{\mathcal{S}, \mathcal{A}, P_0, P, r\}$ denote an MDP with a finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, initial state distribution $P_0 : \Delta(\mathcal{S})$, transition probability $P : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ and reward distribution $r : \mathcal{S} \times \mathcal{A} \mapsto \Delta([0, 1])$, where $\Delta(\mathcal{S})$ denotes a distribution over $\mathcal{S}$. An MDP M and an algorithm $\mathcal{C}$ operating on M with initial state s constitute a stochastic process described by the states s_t visited at time step t, the actions at chosen by $\mathcal{C}$ at step t, and the rewards r_t obtained for $t = 1, \ldots, T$.

A policy on an MDP M is a mapping $\pi : \mathcal{S} \mapsto \mathcal{A}$. We only consider unichain MDPs, in which under any policy, any state can be reached (after a finite number of transitions) from any state. Any policy π on an unichain M generates an ergodic Markov chain with a stationary distribution d_π. An average reward of a policy is defined as

$$\lambda(M, \pi) = \sum_{s \in S} d_\pi(s) r(s, \pi(s)).$$

Let $\pi^*(M)$ denote an optimal policy from $\arg\max_\pi \lambda(M, \pi)$ and let $\lambda^*(M)$ denote the corresponding average reward. Without further specification, π^* and λ^* are those for true MDP M. We call an policy π ϵ-optimal, if $\lambda(M, \pi) > \lambda^* - \epsilon$. The aim of an algorithm is to find the optimal policy by balancing exploring new state-action pairs and exploiting existed good policies. We evaluate an algorithm by total regret defined as $R_T = \sum_{t=0}^{T} \lambda^* - r_t$.

The following definitions are also needed to demonstrate our results.

Definition 2.1. Given an ergodic Markov chain C, let $T_{s,s'} = \{t > 0 \mid s_t = s', s_0 = s\}$ be the first passage time for two states s, s'. Furthermore $T_{s,s}$ is the return time for s. Let $T_C := \max_{s,s' \in S} \mathbb{E}(T_{s,s'})$ and, $\kappa_C := \max_{s \in S} \frac{\max_{s' \neq s} \mathbb{E}(T_{s',s})}{2\mathbb{E}(T_{s,s})}$. Then the hitting time of a unichain MDP M is $T_M := \max_\pi T_{C_\pi}$, where C_π is the Markov chain induced by π on M. Furthermore, we set $\kappa_M := \max_\pi \kappa_{C_\pi}$.

Definition 2.2. Let $\epsilon_M = \lambda^*(M) - \max_{\pi \neq \pi^*} \lambda(M, \pi)$ be the gap between the average rewards of optimal policy and sub-optimal policy for M.

3 Algorithm

Most previous algorithms view the problem as achieving optimism in face of uncertainty (OFU). This class of algorithms maintains a confidence set of MDPs containing the true MDP with high probability. At start of each episode, the algorithm chooses an MDP with an optimal average reward higher than that of true MDP. Instead, we use a similar framework but consider an direction-consistent perturbation (Alg. 1) that perturbs transition probability of all the state-pairs toward the same direction. This perturbation achieves optimism with a probability of $1/S$.

The direction-consistent perturbation is first proposed in Agrawal and Jia (2017), in which it is called optimistic sampling. The paper uses optimistic sampling as a pre-stage. They optimize average rewards among $\tilde{O}(S)$ MDPs from optimistic sampling, which makes it analogous to the previous OFU algorithms. We follow the same sampling idea but only sample once and give a worst case bound on total regret. To ensure that MDP after perturbation is still unichain, we add a small uniform probability to all directions.

Algorithm 1 Direction-consistent perturbation

Input: $N_k(s,a), N_k(s^{'},s,a)$ and $\hat{P}_k(s^{'} \mid s,a)$ for any $s, s^{'} \in \mathcal{S}$ and $a \in \mathcal{A}$; current time step t_k; $\rho \in (0,1)$.

Let $P_k^-(\cdot \mid s,a) = \hat{P}_k(\cdot \mid s,a) - \Delta_k(\cdot \mid s,a)$, where

$$\Delta_k(\cdot \mid s,a) = \min\{\sqrt{\frac{3\hat{P}_k(i \mid s,a)\log(2ASt_k/\rho)}{N_k(s,a)}} + \frac{3\log(2ASt_k/\rho)}{N_k(s,a)}, \hat{P}_k(i \mid s,a)\}, i \in \mathcal{S}. \quad (1)$$

Let $\delta_k = \min_{s,a}\sqrt{\frac{3S\log(2ASt_k/\rho)}{N_k(s,a)}}$. Let z_k be a random vector sampled uniformly at random from $\{\mathbb{1}_1, \dots \mathbb{1}_S\}$, where $\mathbb{1}_i$ is the vector with 1 on dimension i and 0's on the other dimensions. Let $\mathbb{1}$ be the vector with all the elements as 1. Set $\tilde{M}_k$ with the transition probability,

$$\tilde{P}_k(\cdot \mid s,a) = P_k^-(\cdot \mid s,a) + (\sum_i \Delta_k(i \mid s,a) - \delta_k)z_k + \frac{\delta_k}{S}\mathbb{1}.$$

Output: $\tilde{M}_k$

Another difference is the length of episodes. Previous methods mainly use doubling trick which ends an episode when visit number of some state and action pair doubles. It can be shown that doubling trick fails in our perturbation scenario. Thus, we consider a fixed number of steps for each episode. We assume $2T_M \leq H$ in our proof. The whole algorithm is shown in Alg. 2.

Algorithm 2 Perturbation algorithm for reinforcement learning (PRL)

procedure PERTURBATION ALGORITHM($\rho \in (0,1), \mathcal{S}, \mathcal{A}$ and H)
Initialization: Set $t \coloneqq 1$, and observe the initial state s_1.
for episodes $k = 1, 2, \dots$ **do**
Initialize episode k:
1. Set the start time of episode k, $t_k \coloneqq t$.
2. For all $s, s^{'} \in \mathcal{S}, a \in \mathcal{A}$, set the state-action counts prior to episode k,

$$N_k(s,a) \coloneqq \#\{\tau < t_k : s_\tau = s, a_\tau = a\},$$
$$N_k(s,a,s') \coloneqq \#\{\tau < t_k : s_\tau = s, a_\tau = a, s_{\tau+1} = s'\}.$$

Compute estimates $\hat{P}_k(s' \mid s,a) \coloneqq \frac{N_k(s,a,s')}{\max\{1, N_k(s,a)\}}$.

Compute and execute policy π_k:
3. Sample a MDP $\tilde{M}_k$ from algorithm 1 and compute its optimal policy π_k.
4. Execute policy π_k for H steps.
end for
end procedure

4 Main results

We give a logarithmic in T, gap-dependent regret bound with polynomial dependence on S, A, T_M, κ_M that is common in bandit literature .

Theorem 4.1. *With probability of at least $1-\rho$, it holds that for any initial state $s \in \mathcal{S}$, any $T > 1$, the regret of **PRL** on M is bounded by*

$$R_T = O(\frac{AS^5T_M\kappa_M^2\log(SAT/\rho)}{\epsilon_M^2}), \quad (2)$$

if $H > 2T_M$ and M is an unichain with finite state and action space.

4.1 Proof of Theorem 4.1

We prove Theorem 4.1 by constructing a perturbation bound for both $\lambda(\tilde{M}_k, \pi_k)$ and $\lambda(\tilde{M}_k, \pi^*)$. If width of both decrease to $\epsilon/2$, π_k is an ϵ-optimal policy. Our proof mainly consists of bounding the regrets caused by the episodes when the two bounds are wider than $\epsilon/2$ and the regrets of failing the confidence bounds.

Perturbation bound for average reward. Cho and Meyer (2001) gives a perturbation bound on the stationary distribution between two Markov chains. We extend it to MDP setting.

Lemma 4.2. *For a given policy π, let $\mathcal{M}(\hat{P}, \pi, \epsilon_p)$ be a set of MDPs defined as*

$$\{M : \|P(\cdot|s, \pi(s)) - \hat{P}(\cdot|s, \pi(s))\|_1 \leq \epsilon_p, \forall s \in \mathcal{S}\}.$$

Then for any two MDPs, M_1 and M_2, $|\lambda(M_1, \pi) - \lambda(M_2, \pi)| \leq S\epsilon_p \kappa_{M_1}$.

We fix $\epsilon = \epsilon_M$ and a policy π an ϵ-known policy at the start of episode k, if for any $s \in \mathcal{S}$,

$$L_k^p(s, \pi(s)) \leq \epsilon_p \coloneqq \frac{\epsilon}{2S\kappa_M}, \text{ where } L_k^p(s, a) \coloneqq \sqrt{\frac{12S\log(2AST/\rho)}{N_k(s, a)}}. \tag{3}$$

A plausible episode satisfies the following two conditions: (1) π_k and π^* are both ϵ-known; (2) both P and $\tilde{P}_k$ are in the plausible set $\cap_{\pi \in \{\pi_k, \pi^*\}} \mathcal{M}(\hat{P}_k, \pi, L_k^p(s, \pi(s)))$. By Lemma 4.2, with the two conditions satisfied, the optimal policy is achieved, i.e. $\pi_k = \pi^*$.

Therefore, our proof mainly consists of bounding the regret when π_k or π^* is not ϵ-known and bounding the episodes when P or $\tilde{P}_k$ is not in the plausible set. Let A_k denote the event that π_k is ϵ-known at the start of k and B_k denote the event that π^* is ϵ-known at the start of k.

Regret due to unknown π_k. Similar to Auer and Ortner (2007), we first bound the regret caused by episodes running under unknown policies, i.e. A_k^c holds. For some $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $m(s, a)$ be the number of the episodes with $L_k^p(s, a) \geq \epsilon_p$. Since the mean passage time between any state s'' and s is upper bounded by T_M, the probability that it takes more than $2T_M$ steps to reach s from s'' is $\frac{1}{2}$ by Markov's inequality. Since we assume $H > 2T_M$, let $\beta = \frac{H/T_M}{\lfloor H/2T_M \rfloor}$ and separate each round i into $\lfloor \frac{H}{\beta T_M} \rfloor$ intervals of length $\geq 2T_M$, in each of which the visiting probability on state s is at least $\frac{1}{2}$. We may lower bound the number of visits $N_{s,a}(n)$ in (s, a) within n such intervals by an application of Chernoff-Hoeffding's inequality:

$$P\{N_{s,a}(n) \geq \frac{n}{2} - \sqrt{n\log(\frac{SA}{\rho})}\} \geq 1 - \frac{\rho}{SA}. \tag{4}$$

If $N_t(s, a) > \frac{12\mathcal{S}\log(2AST/\rho)}{\epsilon_p^2}$, (3) holds for every step. Combined with (4), we have

$$\sum_{i=1}^{m(s,a)} \lfloor \frac{H}{\beta T_M} \rfloor = O(\frac{S\log(2SAT/\rho)}{\epsilon_p^2}), \tag{5}$$

with probability $1 - \frac{\rho}{SA}$. Summing over (s, a), with probability $1 - \rho$, we have

$$\sum_{k=1}^{T/K} \mathbb{1}_{A_k^c} H \leq \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \sum_{i=1}^{m(s,a)} H = O(\frac{AS^2 T_M \log(2SAT/\rho)}{\epsilon_p^2}) =: N_{A^c}. \tag{6}$$

It remains to bound the regrets caused by episodes with $A_k \cap B_k^c$.

Regret due to unknown π^*. We first introduce the lemma on the probability of optimism under direction-consistent perturbation, whose proof is shown in Appendix A.

Lemma 4.3. *(Agrawal and Jia, 2017) In each episode k, with probability $\frac{1}{2S}$, $\exists i_k \in \mathcal{S}$, when $z_k = i_k$, $\lambda(\tilde{M}_k, \pi^*) > \lambda(M, \pi^*) - \delta_k S^2 \kappa_M$. If π_k is ϵ-known, $\lambda(\tilde{M}_k, \pi^*) > \lambda(M, \pi^*) - \epsilon/2$.*

Let E_k be the event that we choose $z_k = i_k$ in episode k. Note that $E_k \cap A_k$ indicates that π_k is π^*, which means π^* is ϵ-known Thus, $E_k \cap A_k \cap B_k^c = \emptyset$ and $A_k \cap B_k^c$ can only happen when E_k^c holds. We have

$$\#(A_k \cap B_k^c) \leq \#(E_k^c \cap B_k^c) =: N_1, \text{ and } \quad N_2 := \#(E_k \cap B_k^c) = \#(A_k^c \cap B_k^c) \leq \#(A_k^c).$$

Given a fixed $N_2, N_1 \sim NB(N_2, 1 - \frac{1}{2S})$ is a negative binomial distribution. By some tail bound of negative binomial distribution (Klar, 2000), we have Lemma 4.4. The proof is shown in Appendix B.

Lemma 4.4. *With probability at least* $1 - \rho$*,*

$$HN_1 < \max\{SN_2H, S^2 \log(1/\rho)H\} = O(SN_{A^c}). \tag{7}$$

Applying this Lemma, the total regret in episodes when $A_k \cap B_k^c$ happens is $O(SN_{A^c})$. With the confidence bound containing both true MDP M and perturbed MDP $\tilde{M}_k$ at the start of each episode, the total regret is $O(SN_{A^c})$.

Deal with failing confidence set Finally, we have to take into account the error probabilities, with which in each round a transition probability is not contained in its confidence interval.

By Weissman et al. (2003),

$$\mathbb{P}\left\{\|\hat{p}(\cdot) - p(\cdot)\|_1 \geq \alpha\right\} \leq (2^m - 2)\exp\left(-\frac{n\alpha^2}{2}\right).$$

Apply this bound between empirical and true distribution on transition probability. Since

$$\sqrt{\frac{2}{n}\log(\frac{2^S AS t^6}{\rho})} \leq \sqrt{\frac{12S}{n}\log(\frac{2ASt}{\rho})},$$

we have

$$\sum_{s,a}\sum_t \mathbb{P}\left\{\|\hat{P}_t(\cdot \mid s,a) - P_t(\cdot \mid s,a)\|_1 \geq \sqrt{\frac{12S}{N_t(s,a)}\log\left(\frac{2ASt}{\rho}\right)}\right\} \leq \sum_{s,a}\sum_t \frac{\rho}{t^6 SA} = O(\rho). \tag{8}$$

For $\tilde{P}_k$ and any policy π, we have as long as $3S\log(2ASt_k/\rho) < N(s,\pi(s))$,

$$\sum_{i=1}^{S}\Delta_k(i \mid s,\pi(s)) - \delta_k \leq \sqrt{\frac{3S\log(2ASt_k/\rho)}{N_k(s,\pi(s))}} + \frac{3S\log(2ASt_k/\rho)}{N_k(s,\pi(s))} \leq 2\sqrt{\frac{3S\log(2At_k/\rho)}{N_k(s,\pi(s))}} \leq L_k^p(s,\pi(s)).$$

It holds when π is ϵ-known. Putting (7), (6) and (8) together, Theorem 4.1 follows.

5 Discussion and open problems

Our analysis shows that a sufficient condition to have a worst-case bound is to have a significant probability of being optimistic in each episode. The probability for direction-consistent perturbation is $1/S$. However, more general perturbation scenarios, for example, posterior sampling, need to be studied. The other promising extension is to consider communicating setting and episodic setting.

A Proof of Lemma 4.3

Let $\tilde{P}_k^1(\cdot \mid s,a) = P_k^-(\cdot \mid s,a) + (\sum_i \Delta_k(i \mid s,a))$ and $\tilde{P}_k^2(\cdot \mid s,a) = P_k^1(\cdot \mid s,a) - \delta_k z_k + \delta_k$. With some abuse of notations, we omit subscriptions for k and fix a state and action pair.

Lemma A.1. *For any fixed* $h \in [0,D]^S$*, we have*

$$\tilde{P}^{1T}h \geq P^T h,$$

with probability at least $\Omega(1/S)$

Define $\alpha_i \coloneqq \hat{P}(i) - P(i)$ (and hence $\sum_i \alpha_i = 0$). By multiplicative Chernoff bounds (Kleinberg et al., 2008), with probability $1 - \frac{\rho}{SAt}$, $|\alpha_i| \leq \sqrt{\frac{3\hat{P}(i)\log(2ASt/\rho)}{N}} + \frac{3\log(2ASt/\rho)}{N}$, and with probability $1 - \frac{\rho}{At}$, it holds for any $i \in \mathcal{S}$. Note that $\Delta_i \geq \alpha_i$. With probability $1/S$, $z = i$ is picked such that $h_i = D$. So with probability at least $\frac{1}{2S}$,

$$\begin{aligned}
\sum_i \tilde{P}^1(i)h(i) &= \sum_i P(i)^- h(i) + D\left(1 - \sum_j P(j)^-\right) = \sum_i P(i)^- h(i) + D\sum_j \Delta(j) \\
&= \sum_i \left(\hat{P}(i) - \Delta(i)\right) h(i) + D\Delta(i) = \sum_i \hat{P}(i)h(i) + (D - h(i))\,\Delta(i) \\
&\geq \sum_i \hat{P}(i)h(i) + (D - h(i))\,\alpha(i) = \sum_i \left(\hat{P}(i) - \alpha(i)\right) h_i + D\alpha(i) \\
&= \sum_i P(i)h(i) + D\sum_i \alpha(i) = \sum_i P(i)h(i)
\end{aligned}$$

By Agrawal and Jia (2017) (Lemma 4.2), $\lambda(\tilde{M}^1, \pi^*) \geq \lambda(M, \pi^*)$ with probability at least $1/2S$. Applying perturbation bound again, $\lambda(\tilde{M}^2, \pi^*) \geq \lambda(M, \pi^*) - \delta S^2 \kappa_M$.

B Proof of Lemma 4.4

Lemma B.1. *By Klar (2000) (Proposition 1), suppose X has a negative binomial distribution with parameters r and p, where $r > 0$, and $0 < p < 1$; f_x is the density function. If $r > 1$ and $n \geq rq/p$,*

$$\frac{1}{p} f_n < P(X \geq n) < \left(1 - \frac{n+r}{n+1} q\right)^{-1} f_n.$$

Applying Lemma B.1, with $c \geq 2S$,

$$\begin{aligned}
\Pr(N_1 \geq cN_2) &\leq [1 - \frac{(1+c)N_2}{cN_2+1}\frac{S}{S+1}]^{-1} \begin{pmatrix} (1+c)N_2 - 1 \\ cN_2 \end{pmatrix} (\frac{1}{S+1})^{N_2} (1 - \frac{1}{S+1})^{cN_2} \\
&\quad \text{(Using } 1 - \frac{(1+c)N_2}{cN_2+1}\frac{S}{S+1} \geq 1 - \frac{1+c}{c}\frac{S}{S+1} \geq \frac{1}{2(S+1)}) \\
&\leq 2(S+1) \begin{pmatrix} (1+c)N_2 - 1 \\ cN_2 \end{pmatrix} (\frac{1}{S+1})^{N_2} (1 - \frac{1}{S+1})^{cN_2} \\
&\leq 2(S+1)(1 - F(N_2, (1+c)N_2, \frac{1}{S+1})),
\end{aligned}$$

where $F(k, n, p)$ is the cdf of binomial distribution of B(n, p). By Hoeffding's inequality, choosing $cN_2 \geq S^2 \log(2(S+1)/\rho)$,

$$\begin{aligned}
2(S+1)F(N_2, (1+c)N_2, \frac{1}{S+1}) &\leq 2(S+1)\exp\left(-2\frac{((1+c)N_2/S - N_2)^2}{(1+c)N_2}\right) \\
&\leq 2(S+1)\exp\left(-2\frac{(1+c)N_2}{S^2}\right) \leq \rho
\end{aligned}$$

Thus, $P(N_1 > \max\{2SN_2, S^2\log(2(S+1)/\rho)\}) \leq 1 - \rho$.

References

Agrawal, S. and Jia, R. (2017). Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194.

Auer, P. and Ortner, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 49–56.

Cho, G. E. and Meyer, C. D. (2001). Comparison of perturbation bounds for the stationary distribution of a markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150.

Dann, C. and Brunskill, E. (2015). Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307.

Kearns, M. and Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232.

Kim, B. and Tewari, A. (2019). On the optimality of perturbations in stochastic and adversarial multi-armed bandit problems. *arXiv preprint arXiv:1902.00610*.

Klar, B. (2000). Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14(2):161–171.

Kleinberg, R., Slivkins, A., and Upfal, E. (2008). Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690. ACM.

Kujala, J. and Elomaa, T. (2005). On following the perturbed leader in the bandit setting. In *International Conference on Algorithmic Learning Theory*, pages 371–385. Springer.

Osband, I., Russo, D., and Van Roy, B. (2013). (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, pages 3003–3011.

Osband, I. and Van Roy, B. (2017). Why is posterior sampling better than optimism for reinforcement learning? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2701–2710. JMLR. org.

Russo, D. (2019). Worst-case regret bounds for exploration via randomized value functions. *arXiv preprint arXiv:1906.02870*.

Van Erven, T., Kotłowski, W., and Warmuth, M. K. (2014). Follow the leader with dropout perturbations. In *Conference on Learning Theory*, pages 949–974.

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.