Active Exploration in Markov Decision Processes

Jean Tarbouriech
Facebook AI Research

Alessandro Lazaric
Facebook AI Research

Abstract

We introduce the active exploration problem in Markov decision processes (MDPs). Each state of the MDP is characterized by a random value and the learner should gather samples to estimate the mean value of each state as accurately as possible. Similarly to active exploration in multi-armed bandit (MAB), states may have different levels of noise, so that the higher the noise, the more samples are needed. As the noise level is initially unknown, we need to trade off the *exploration* of the environment to estimate the noise and the *exploitation* of these estimates to compute a policy maximizing the accuracy of the mean predictions. We introduce a novel learning algorithm to solve this problem showing that active exploration in MDPs may be significantly more difficult than in MAB. We also derive a heuristic procedure to mitigate the negative effect of slowly mixing policies. Finally, we validate our findings on simple numerical simulations.

1 Introduction

Active exploration[1] refers to the problem of actively querying an unknown environment to gather information and perform accurate predictions about its behavior. Popular instances of active exploration are optimal design of experiments (Pukelsheim, 2006) and, more in general, active learning (AL) (Hanneke, 2014), where given a fixed budget of samples, a learner actively chooses where to query an unknown function to collect information that could maximize the accuracy of its predictions. An effective AL method should adjust to the *approximation function space* to obtain samples wherever the uncertainty is high. In multi-armed bandit (MAB), the active exploration problem (Antos et al., 2010; Carpentier et al., 2011) rather focuses on adjusting to the *noise* affecting the observations, which may differ over arms. Despite their difference, in both AL and MAB, the underlying assumption is that the learner can *directly* collect a sample at any arbitrary point or pull any arm with no constraint.

In this paper, we extend the MAB setting to active exploration in a Markov decision process (MDP), where each state (an arm in the MAB setting) is characterized by a random variable that we need to estimate. Unlike AL and MAB, if the learner needs to generate an "experiment" at a state, it needs to move from the current state to the desired state. Consider the problem of accurately measuring the level of pollution over different sites when a fixed budget of measurements is provided and only one measuring station is available. The noise affecting the observations may differ over sites and we need to carefully design a policy in order to collect more samples (resp. less samples) on sites with higher (resp. lower) noise. Since the noise level may be unknown in advance, this requires alternating between the exploration of the environment to estimate the noise level and the exploitation of the estimates to optimize the collection of "useful" samples.

The main contributions of this paper can be summarized as follows: **1)** we introduce the active exploration problem in MDP and provide a thorough discussion on its difference w.r.t. the bandit case, **2)** inspired by the bandit algorithm of Carpentier et al. (2011) and Frank-Wolfe UCB by Berthet and Perchet (2017), we devise a novel learning algorithm with vanishing regret under the assumption that the MDP is ergodic and its dynamics is known in advance, **3)** we discuss how slowly mixing policies may compromise the estimation accuracy and introduce a heuristic convex problem to compute faster mixing reversible policies, **4)** we report numerical simulations on simple MDPs to validate our theoretical findings. Finally, we discuss how our as-

[1] We use this term in contrast to the exploration-exploitation dilemma (i.e., regret minimization) and best-arm identification.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

sumptions (e.g., known dynamics) could be relaxed.

Related work. Dance and Silander (2017) study active exploration in restless bandit where the value of each arm is not an i.i.d. random variable but has a stationary dynamics. Nonetheless, they still consider the case where any arm can be pulled at each time step. Security games, notably the patrolling problem (e.g., Basilico et al., 2012), often consider the dynamics of moving from a state to another, but the active exploration is designed to contrast an adversary "attacking" a state (e.g., Balcan et al., 2015). Rolf et al. (2018) consider the problem of navigating a robot in an environment with background emissions to identify the k strongest emitters. While the performance depends on how the robot traverses the environment, the authors only consider a fixed sensing path. Auer et al. (2011) study the autonomous exploration problem, where the objective is to discover the set of states that are reachable (following a shortest path policy) within a given number of steps. Intrinsically motivated reinforcement learning (Chentanez et al., 2005) often tackles the problem of "discovering" how the environment behaves (e.g., its dynamics) by introducing an "internal" reward signal. Hazan et al. (2018) recently focus on the instrinsically-defined objective of learning a (possibly non-stationary) policy that induces a state distribution that is as uniform as possible (i.e., with maximal entropy). This problem is related to our setting in the special case of equal state variances. We believe such line of work is insightful as it may help to understand how to encourage an agent to find policies which can manipulate its environment in the absence of any extrinsic scalar reward signal.

2 Preliminaries

Active exploration in MDPs. A Markov decision process (MDP) (Puterman, 1994) is a tuple $M = (\mathcal{S}, \mathcal{A}, p, \nu, \overline{s})$, where $\mathcal{S}$ is a set of S states, $\mathcal{A}$ is a set of A actions, and for any $s, a \in \mathcal{S} \times \mathcal{A}$, $p(s'|s,a)$ is the transition distribution over the next state $s' \in \mathcal{S}$. We also define the adjacency matrix $Q \in \mathbb{R}^{S \times S}$, such that $Q(s,s') = 1$ for any $s, s' \in \mathcal{S}$ where there exists an action $a \in \mathcal{A}$ with $p(s'|s,a) > 0$, and $Q(s,s') = 0$ otherwise. Instead of a reward function, $\nu(s)$ is an observation distribution supported in $[0, R]$ with mean $\mu(s)$ and variance $\sigma^2(s)$, characterizing the random event that we want to accurately estimate on each state. Finally, $\overline{s}$ is the starting state. The stochastic process works as follows. At step $t = 1$ the environment is initialized at $s_1 = \overline{s}$, an agent takes an action a_1, which triggers a transition to the next state $s_2 \sim p(\cdot|s_1, a_1)$ and an observation $x_2 \sim \nu(s_2)$, and so on. We denote by $\mathcal{F}_t = \{s_1, a_1, s_2, x_2, a_2, \dots, s_t, x_t\}$ the history up to t. A randomized history-dependent (resp. stationary) policy π at time t is denoted by $\pi_t : \mathcal{F}_t \to \Delta(\mathcal{A})$ (resp. $\pi : \mathcal{S} \to \Delta(\mathcal{A})$) and it maps the history (resp. the current state) to a distribution over actions. We denote the set of history-dependent (resp. stationary) policies by Π^{HR} (resp. Π^{SR}). For any policy π, we denote by $T_{\pi,n}(s) = \sum_{t=2}^{n} \mathbb{I}\{s_t = s\}$ the number of observations collected in state s when starting from $s_1 = \overline{s}$ and following policy π for n steps.[2] At the beginning of step t, for any state s such that $T_{\pi,t}(s) > 0$, the empirical estimates of the mean and variance are computed as

$$\begin{aligned} \widehat{\mu}_{\pi,t}(s) &= \frac{1}{T_{\pi,t}(s)} \sum_{\tau=2}^{t} x_\tau \mathbb{I}\{s_\tau = s\} \\ \widehat{\sigma}^2_{\pi,t}(s) &= \frac{1}{T_{\pi,t}(s)} \sum_{\tau=2}^{t} x_\tau^2 \mathbb{I}\{s_\tau = s\} - \widehat{\mu}^2_{\pi,t}(s) \end{aligned} . \quad (1)$$

In order to avoid dealing with subtle limit cases and simplify the definition of the estimation problem, we introduce the following assumption.

Assumption 1. *For any state $s \in \mathcal{S}$ and policy π, $T_{\pi,1}(s) = 1$ and $T_{\pi,n}(s) = 1 + \sum_{t=2}^{n} \mathbb{I}\{s_t = s\}$.*

We basically assume that at $t = 1$ one sample is available and used in estimating $\mu(s)$ and $\sigma^2(s)$ at each state (see App. A.1 for further discussion). For any policy π and any budget $n \in \mathbb{N}$, we define the estimation problem as the minimization of the normalized mean-squared estimation error

$$\min_{\pi \in \Pi^{\mathrm{HR}}} \mathcal{L}_n(\pi) := \frac{n}{S} \sum_{s \in \mathcal{S}} \mathbb{E}_{\pi,\nu}\Big[\big(\widehat{\mu}_{\pi,n}(s) - \mu(s)\big)^2\Big],$$

where $\mathbb{E}_{\pi,\nu}$ is the expectation w.r.t. the trajectories generated by π and the observations from ν. When the dynamics p and the variances $\sigma^2(s)$ are known, we restrict our attention to stationary polices $\pi \in \Pi^{\mathrm{SR}}$ and exploiting the independence between transitions and observations, and Asm. 1, we obtain

$$\begin{aligned} \mathcal{L}_n(\pi) &= \frac{n}{S} \sum_{s \in \mathcal{S}} \mathbb{E}_\pi\Big[\mathbb{E}_\nu\Big[\big(\widehat{\mu}_{\pi,n}(s) - \mu(s)\big)^2 \Big| T_{\pi,n}\Big]\Big] \\ &= \frac{1}{S} \sum_{s \in \mathcal{S}} \sigma^2(s) \mathbb{E}_\pi\left[\frac{n}{T_{\pi,n}(s)}\right]. \end{aligned} \quad (2)$$

In the case of deterministic and fully-connected MDPs, the problem smoothly reduces to the active bandit formulation of Antos et al. (2010).

Technical tools. For any stationary policy $\pi \in \Pi^{\mathrm{SR}}$, we denote by P_π the kernel of the Markov chain induced by π in the MDP, i.e., $P_\pi(s'|s) = \sum_{a \in \mathcal{A}} p(s'|s,a)\pi(a|s)$. If the Markov chain P_π is ergodic (i.e., all states are aperiodic and recurrent), it admits a unique stationary distribution over states

[2] The counter starts at 2 as observations are received upon arrival on a state (i.e., no observation at $s_1 = \overline{s}$).

η_π, such that $\eta_\pi(s) = \sum_{s'} P_\pi(s|s')\eta_\pi(s')$. A Markov chain P_π is reversible if the detailed balance condition $\eta_\pi(s)P_\pi(s'|s) = \eta_\pi(s')P_\pi(s|s')$ is satisfied for all $s, s' \in \mathcal{S}$. Let $\{\xi_\pi(s)\}$ be the eigenvalues of P_π, we define the second-largest eigenvalue modulus (SLEM) and the spectral gap as

$$\xi_{\pi,\max} := \max_{s:\xi_\pi(s)\neq 1} |\xi_\pi(s)|; \quad \gamma_\pi := 1 - \xi_{\pi,\max}. \tag{3}$$

The SLEM can be written as the spectral norm (i.e., the maximum singular value) of an affine matrix in P_π. Let D_η be the diagonal matrix with the elements of η, then (Boyd et al., 2004)

$$\xi_{\pi,\max} = \|D_{\eta_\pi}^{1/2} P_\pi D_{\eta_\pi}^{-1/2} - \sqrt{\eta_\pi}\sqrt{\eta_\pi}^T\|_2. \tag{4}$$

For ergodic chains, $\xi_{\pi,\max} < 1$. The spectral gap is tightly related to the mixing time of the chain and it characterizes how fast the frequency of visits converges to the stationary distribution (e.g., Hsu et al. (Thm. 3, 2015), Paulin et al. (Thm. 3.8, 2015)).

Proposition 1. *Let $\pi \in \Pi^{SR}$ be a stationary policy inducing an ergodic and reversible chain P_π with spectral gap γ_π and stationary distribution η_π. Let $\eta_{\pi,\min} = \min_{s\in\mathcal{S}} \eta_\pi(s)$. For any budget $n > 0$ and state $s \in \mathcal{S}$,*

$$\Big|\frac{\mathbb{E}\big[T_{\pi,n}(s)\big]}{n} - \eta_\pi(s)\Big| \le \frac{1}{2\sqrt{\eta_{\pi,\min}}}\frac{1}{\gamma_\pi n},$$

and for any $\delta \in (0,1)$, with probability $1-\delta$,

$$\Big|\frac{T_{\pi,n}(s)}{n} - \eta_\pi(s)\Big| \le \epsilon_\pi(s,n,\delta) := O\Big(\sqrt{\frac{\ln(\frac{1}{\delta}\sqrt{\frac{2}{\eta_{\pi,\min}}})}{\gamma_\pi n}}\Big).$$

The exact formulation of $\epsilon_\pi(s,n,\delta)$ is reported in App. C (see proof of Lem. 1). It is interesting to notice that the convergence in expectation is faster than in high-probability ($O(n^{-1})$ vs $O(n^{-1/2})$), but in both cases the spectral gap may significantly affect the convergence (e.g., for slowly mixing chains $\gamma_\pi \approx 0$).

Finally, we recall a concentration inequality for variance estimation (see Antos et al. (2010)).

Proposition 2. *For any $\delta \in (0,1)$ and time t, with probability at least $1-\delta$*

$$\big|\widehat{\sigma}_t^2(s) - \sigma^2(s)\big| \le \alpha(t,s,\delta) := 5R^2\sqrt{\frac{\log(\frac{4St}{\delta})}{T_t(s)}}.$$

3 The Asymptotic Case

In deterministic fully-connected MDPs, problem (2) reduces to the bandit setting and it also inherits its NP-hard complexity, as it may require enumerating all possible values of $\{T_n(s)\}_s$ (see e.g., Welch, 1982). In our case, this difficulty is further increased by the fact that observations can only be collected through the "constraint" of the MDP dynamics. In this section we introduce an asymptotic version of the estimation problem and a learning algorithm with vanishing regret w.r.t. the optimal asymptotic stationary policy.

3.1 An Asymptotic Formulation

A standard approach in experimental optimal design (Pukelsheim, 2006) and MAB (Antos et al., 2010; Carpentier et al., 2011) is to replace problem (2) by its continuous relaxation, where the empirical frequency $T_n(s)/n$ is replaced by a distribution over states. In our case $T_n(s)$ cannot be directly selected so we rather consider an asymptotic formulation for $n \to \infty$.[3] We first introduce the following assumption on the MDP.

Assumption 2. *For any stationary policy $\pi \in \Pi^{SR}$, the corresponding Markov chain P_π is ergodic and we denote by $\eta_{\min} = \inf_{\pi\in\Pi^{SR}} \min_{s\in\mathcal{S}} \eta_\pi(s)$ the smallest stationary probability across policies.*

Asm. 1 and 2, together with the continuity of the inverse function $x \mapsto 1/x$ on $[1/n, 1]$, guarantee that for any policy π, $\frac{n}{T_{\pi,n}(s)}$ converges almost-surely to $\frac{1}{\eta_\pi(s)}$ when $n \to +\infty$ (see Prop. 1). As a result, we replace problem (2) with

$$\begin{aligned} &\min_{\pi\in\Pi^{\mathrm{SR}},\eta\in\Delta(\mathcal{S})} \mathcal{L}(\pi,\eta) := \frac{1}{S}\sum_{s\in\mathcal{S}} \frac{\sigma^2(s)}{\eta(s)} \\ &\text{s.t. } \forall s \in \mathcal{S},\ \eta(s) = \sum_{s',a} \pi(a|s')p(s|s',a)\eta(s') \end{aligned}, \tag{5}$$

where η is constrained to be the stationary distribution associated with π (i.e., $\eta = \eta_\pi$). While both π and η belong to a convex set and $\mathcal{L}(\pi,\eta)$ is convex in η, the overall problem is not convex because of the constraint. Yet, we can apply the same reparameterization used in the dual formulation of reward-based MDP (Sect. 8, Puterman, 1994) and introduce the state-action stationary distribution $\lambda_\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ of a policy π. Let

$$\begin{aligned} \Lambda = \Big\{\lambda \in \Delta(\mathcal{S}\times\mathcal{A}) :\ &\forall s \in \mathcal{S}, \\ &\sum_{b\in\mathcal{A}} \lambda(s,b) = \sum_{s'\in\mathcal{S},a\in\mathcal{A}} p(s|s',a)\lambda(s',a)\Big\} \end{aligned} \tag{6}$$

be the set of state-action stationary distributions, we define the optimization problem

$$\begin{aligned} &\min_{\lambda\in\Delta(\mathcal{S}\times\mathcal{A})} \mathcal{L}(\lambda) := \frac{1}{S}\sum_{s\in\mathcal{S}} \frac{\sigma^2(s)}{\sum_{a\in\mathcal{A}}\lambda(s,a)}. \\ &\text{subject to} \quad \lambda \in \Lambda \end{aligned} \tag{7}$$

We can characterize this problem as follows.

[3] In the bandit case, the continuous relaxation is equivalent to the asymptotic formulation.

Proposition 3. *The function $\mathcal{L}(\lambda)$ is convex on the convex set Λ. Let $\lambda^\star$ be the solution of (7), then the policy*

$$\pi_{\lambda^\star}(a|s) = \frac{\lambda^\star(s,a)}{\sum_{b\in\mathcal{A}} \lambda^\star(s,b)},\ \forall s \in \mathcal{S}, a \in \mathcal{A} \quad (8)$$

belongs to Π^{SR} and solves problem (5). Furthermore for any $\underline{\eta} > 0$, $\mathcal{L}(\lambda)$ is $C_{\underline{\eta}}$-smooth on the restricted set $\Lambda_{\underline{\eta}} = \{\lambda \in \Lambda : \sum_{a\in\mathcal{A}} \lambda(s,a) \geq 2\underline{\eta},\ \forall s \in \mathcal{S}\}$ with parameter $C_{\underline{\eta}} \leq A \sum_{s\in\mathcal{S}} \sigma^2(s)/(2\underline{\eta})^3$.

As a result, whenever the dynamics of the MDP and the variances $\sigma^2(s)$ are known, problem (7) can be efficiently solved using any optimization algorithm for convex and smooth functions (e.g., projected gradient descent or Frank-Wolfe (Jaggi, 2013)). Leveraging Prop. 1 we can also characterize the difference between the solutions of the asymptotic problem (5) and the finite-budget one (2). For the sake of simplicity and at the cost of generality (see App. A.2), we introduce an additional assumption.

Assumption 3. *For any stationary policy $\pi \in \Pi^{SR}$, the corresponding Markov chain P_π is reversible and we denote by $\gamma_{\min} = \min_{\pi\in\Pi^{SR}} \gamma_\pi$ the smallest spectral gap across all policies.*

Lemma 1. *Let $\delta = SA^S/n^2$, if n is big enough such that for any $s \in \mathcal{S}$ and any stationary policy $\pi \in \Pi^{SR}$, $\epsilon_\pi(s,n,\delta) \leq \eta_\pi(s)/2$, then we have*

$$\left|\mathcal{L}_n(\pi) - \mathcal{L}(\pi, \eta_\pi)\right| \leq \ell_n(\pi), \quad (9)$$

where

$$\ell_n(\pi) := \frac{1}{S\sqrt{\eta_{\min}} n \gamma_\pi} \sum_{s\in\mathcal{S}} \frac{\sigma^2(s)}{\eta_\pi^2(s)} \Big(1 + 2\frac{\epsilon_\pi(s,n,\delta)}{\eta_\pi(s)}\Big),$$

which gives the performance loss

$$\mathcal{L}_n(\pi_{\lambda^\star}) - \mathcal{L}_n(\pi_n^\star) \leq \ell_n(\pi_{\lambda^\star}) + \ell_n(\pi_n^\star), \quad (10)$$

where $\pi_n^\star$ is the solution to problem (2) and $\pi_{\lambda^\star}$ is defined in (8).

It is interesting to compare the result above to the bandit case. For $n \geq 4/(S\eta_{\min}^2)$ the performance loss of the continuous relaxation in bandit is bounded as $8\sigma_{\max}^2/(\eta_{\min}^3 n^2)$ (see Prop. 7 in App. C). While the condition on n in Lem. 1 is similar (i.e., from the definition of $\epsilon_\pi(s,n,\delta)$, we need $n > \widetilde{\Omega}(1/\eta_{\min}^2)$), the performance loss differs over two main elements: *(i)* the rate of convergence in n, *(ii)* the presence of the spectral gap γ_π. In MAB, the "fast" convergence rate is obtained by exploiting the smoothness of the function $\mathcal{L}$, which characterizes the performance of both discrete and continuous allocations. On the other hand, in the MDP case, while $\mathcal{L}$ is indeed smooth on the restricted simplex, $\mathcal{L}_n$ is a more complicated function of π, which does not allow the same proof technique to be directly applied. Furthermore, the spectral gap directly influences the difference between the finite-time and asymptotic behavior of a policy π. This extra "cost" is not present in MAB, where any allocation over states can be directly "executed" without waiting for the policy to mix.

Algorithm 1 FW-AME: the *Frank-Wolfe for Active MDP Exploration* algorithm

Input: $\widetilde{\lambda}_1 = 1/SA$, $\underline{\eta}$
for $k = 1, 2, ..., K-1$ **do**
  $\widehat{\psi}_{k+1}^+ = \operatorname{argmin}_{\lambda\in\Lambda_{\underline{\eta}}} \langle \nabla \widehat{\mathcal{L}}_{t_k-1}^+(\widetilde{\lambda}_k), \lambda\rangle$
  $\widehat{\pi}_{k+1}^+(a|s) = \dfrac{\widehat{\psi}_{k+1}^+(s,a)}{\sum_{b\in\mathcal{A}} \widehat{\psi}_{k+1}^+(s,b)}$
  Execute $\widehat{\pi}_{k+1}^+$ for τ_k steps
  Update the state-action frequency $\widetilde{\lambda}_{k+1}$
end for

3.2 Learning Algorithm

We introduce a learning algorithm to incrementally solve the active exploration problem in the setting where the state variances $\sigma^2(s)$ are unknown. We rely on the following assumption.

Assumption 4. *The MDP model p is known.*

In App. A.3 we sketch a way to relax Asm. 4 by following an optimistic approach similar to UCRL (Jaksch et al., 2010) in order to incorporate the uncertainty on the MDP dynamics, and we conjecture that the regret guarantees of the algorithm would remain unchanged.

Let $\underline{\eta} < 1/(2S)$ be a positive constant. Since $\mathcal{L}(\lambda)$ is smooth in $\Lambda_{\underline{\eta}}$ (Prop. 3), it can be optimized using the Frank-Wolfe (FW) algorithm (Jaggi, 2013), which constructs a sequence of linear optimization problems whose solutions are used to incrementally update the candidate solution to problem (7). In MAB, Berthet and Perchet (2017) showed that FW can be fed with optimistic estimates of the gradient to obtain a bandit algorithm with small regret. The resulting algorithm (Frank-Wolfe-UCB) actually reduces to the algorithm of Carpentier et al. (2011) when the function to optimize is the mean estimation error. The mapping from FW to a bandit algorithm relies on the fact that the solution to the linear problem at each iteration of FW corresponds to selecting one single arm. Unfortunately, in the MDP case, FW returns a state-action stationary distribution, which cannot be directly "executed". We then need to adapt the bandit-FW idea to *track* the (optimistic) FW solutions.

In Alg. 1 we illustrate FW-AME (FW for Active MDP Exploration) which proceeds through episodes and is evaluated according to the frequency of visits of each state, i.e., $\widetilde{\lambda}_k(s) = T_{t_k-1}(s)/(t_k-1)$. At the beginning of episode k, FW-AME solves an MDP with reward related to the current estimation error, so that the corresponding optimal policy tends to explore states where the current estimate of the mean $\mu(s)$ is not accurate enough. More formally, FW-AME solves a linear problem to compute the state-action stationary distribution $\widehat{\psi}^+_{k+1}$ minimizing the expected "optimistic" gradient evaluated at the current solution obtained using the confidence intervals of Prop. 2, i.e.,

$$\nabla \widehat{\mathcal{L}}^+_{t_k-1}(\lambda)(s,a) = -\frac{\widehat{\sigma}^2_{t_k-1}(s) + \alpha(t_k-1,s,\delta)}{(\sum_b \lambda(s,b))^2}.$$

This choice favors exploration towards states whose loss is possibly high (i.e., large gradient) and poorly estimated (large confidence intervals). This step effectively corresponds to solving an MDP with a reward equal to $\nabla \widehat{\mathcal{L}}^+_{t_k-1}$. Then the policy $\widehat{\pi}^+_{k+1}$ associated to $\widehat{\psi}^+_{k+1}$ is executed for τ_k steps and the solution $\widetilde{\lambda}_k$ is updated accordingly. Let $\nu_{k+1}(s,a)$ be the number of times action a is taken at state s during episode k. We can write the update rule for the candidate solution as

$$\begin{aligned}\widetilde{\lambda}_{k+1} &= \frac{\tau_k}{t_{k+1}-1}\widetilde{\psi}_{k+1} + \frac{t_k-1}{t_{k+1}-1}\widetilde{\lambda}_k \\ &= \beta_k \widetilde{\psi}_{k+1} + (1-\beta_k)\widetilde{\lambda}_k,\end{aligned}$$

where $\widetilde{\psi}_{k+1}(s,a) = \nu_{k+1}(s,a)/\tau_k$ is the frequency of visits within episode k and $\beta_k = \tau_k/(t_{k+1}-1)$ is the weight (or learning rate) used in updating the solution. While we conjecture that a similar approach could be paired with other optimization algorithms (e.g., projected gradient descent), by building on FW we obtain a projection-free algorithm, where at each episodes we only need to solve a specific instance of an MDP. In App. D we derive the following regret guarantee.

Theorem 1. *Let episode lengths satisfy $t_k = \tau_1(k-1)^3 + 1$ where τ_1 is the length of the first episode, i.e.,*

$$\tau_k = \tau_1(3k^2 - 3k + 1) \quad \textit{and} \quad \beta_k = \frac{3k^2-3k+1}{k^3}.$$

Under Asm. 1, 2, 3, 4, FW-AME *satisfies with high probability*[4]

$$\mathcal{L}(\widetilde{\lambda}_K) - \mathcal{L}(\lambda^\star) = \widetilde{O}\big(t_K^{-1/3}\big).$$

Sketch of the proof. The proof combines the FW analysis, the contribution of the estimated optimistic gradient, and the gap between the target distribution $\widehat{\psi}^+_{k+1}$ and the empirical frequency $\widetilde{\psi}_{k+1}$. Let $\rho_{k+1} := \mathcal{L}(\widetilde{\lambda}_{k+1}) - \mathcal{L}(\lambda^\star)$ be the regret at the end of episode k. Introducing $\psi^\star_{k+1} := \operatorname{argmin}_{\lambda \in \Lambda_{\underline{\eta}}} \langle \nabla \mathcal{L}(\widetilde{\lambda}_k), \lambda \rangle$ and exploiting the convexity and $C_{\underline{\eta}}$-smoothness of $\mathcal{L}$, it is possible to obtain the "recursive" inequality

$$\rho_{k+1} \le (1-\beta_k)\rho_k + C_{\underline{\eta}}\beta_k^2 + \beta_k \epsilon_{k+1} + \beta_k \Delta_{k+1},$$

where $\epsilon_{k+1} := \langle \nabla \mathcal{L}(\widetilde{\lambda}_k), \widehat{\psi}^+_{k+1} - \psi^\star_{k+1} \rangle$ and $\Delta_{k+1} := \langle \nabla \mathcal{L}(\widetilde{\lambda}_k), \widetilde{\psi}_{k+1} - \widehat{\psi}^+_{k+1} \rangle$. The term ϵ_{k+1} is an *optimization error* and it can be effectively bounded exploiting the fact that $\widehat{\psi}^+_{k+1}$ is the result of an optimistic optimization. On the other hand, the term Δ_{k+1} is a *tracking error* and it can be only bounded using Prop. 1 as $1/\sqrt{\tau_k}$. Solving the recursion for the specific choice of t_k in the theorem provides the final bound.

Remark (rate). The most striking difference between this bound and the result of Carpentier et al. (2011) and Antos et al. (2010) in MAB is the worse rate of convergence, $O(t^{-1/3})$ vs $O(t^{-1/2})$. This gap is the result of trading off the "optimization" convergence speed of FW and the tracking performance obtained by executing $\widehat{\pi}^+_{k+1}$. Berthet and Perchet (2017) show that in MAB, the learning rate β_k is set to $1/t$ (as in standard FW) to achieve a $O(t^{-1/2})$ convergence rate. In our case, we can obtain such learning rate by setting episodes of constant length $\tau_k = \tau$. Unfortunately, this scheme would suffer a constant regret. In fact, while a FW instance where the solution $\widetilde{\lambda}_k$ is updated directly using $\widehat{\psi}^+_{k+1}$ would indeed converge faster with such episode scheme, our algorithm cannot "play" the distribution $\widehat{\psi}^+_{k+1}$ but needs to execute the corresponding policy $\widehat{\pi}^+_{k+1}$, which gathers samples with frequency $\widetilde{\psi}_{k+1}$, then used to update $\widetilde{\lambda}_k$. The gap between $\widehat{\psi}^+_{k+1}$ and $\widetilde{\psi}_{k+1}$ reduces the efficiency of the optimization step by introducing an additive error of order $O(1/\sqrt{\tau})$ (see Prop. 1), which is constant for fixed-sized episodes. As a result, the episode length is optimized to trade off between the optimization speed and tracking effectiveness. Whether this is an intrinsic issue of the active exploration in MDP or better algorithms can be devised is an open question.

Remark (problem-dependent constants). Investigating the proof reveals a number of other dependencies on the algorithm's and problem's parameters. First, the regret bound depends on the inverse of the parameter $\underline{\eta}$ used in FW-AME to guarantee the smoothness of the function. While this may suggest to take $\underline{\eta}$ as large as possible, this may over-constrain the optimization problem (i.e., the set $\Lambda_{\underline{\eta}}$ becomes artificially too small). If $\lambda^\star$ is the solution on the "unconstrained" Λ, then $2\underline{\eta}$ should be set exactly at $\min_s \sum_a \lambda^\star(s,a)$. Furthermore, the gap between $\widehat{\psi}^+_{k+1}$

[4] See App. D.2 for a more explicit bound. See App. A for a discussion on the relaxation of Asm. 1, 3 and 4.

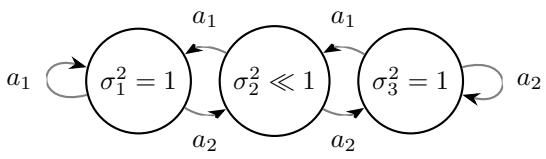


Figure 1: Deterministic 3-state 2-action MDP with $\sigma_1^2 = \sigma_3^2 = 1$ and $\sigma_2^2 \ll 1$.

and $\widetilde{\psi}_{k+1}$ is bounded using Prop. 1. Since the policy executed at each step is random (it depends on the samples observed at previous episodes), we need to take the worst case w.r.t. all possible stationary policies. Thus the regret presents an inverse dependency on $\gamma_{\min}$, which could be very small. Finally, the bound has a direct dependency on the number of states.

4 The Mixing Issue

When the budget n is small, the solution of (7) may be very inefficient compared to the optimal finite-time policy. As an illustrative example, consider the MDP in Fig. 1. In the "unconstrained" version of the problem, where states can be directly sampled (i.e., the bandit setting), the optimal continuous allocation for problem (2) tends to $(0.5, 0, 0.5)$ as $\sigma^2(s_2)$ tends to 0. As soon as we introduce the constraint of the MDP structure, such allocation may not be realizable by any policy. In this MDP, solving problem (7) returns a policy that executes the self-loop actions in s_1 and s_3 with high probability (thus moving to s_2 with low probability) and takes a uniformly random action in s_2. The resulting asymptotic performance does indeed approach the optimal unconstrained allocation, as the stationary distribution of the policy $(\eta(s_1), \eta(s_2), \eta(s_3))$ tends to $(0.5, 0, 0.5)$ for any arbitrary initial state $\overline{s}$. However for any finite budget n, this policy performs very poorly since the agent would get stuck in s_1 (or s_3 depending on the initial state) almost indefinitely, thus making the mean estimation of s_3 (or s_1) arbitrarily bad. As a result, the optimal asymptotic policy mixes arbitrarily slowly as $\sigma^2(s_2)$ tends to zero and its finite-time performance is then arbitrarily far from the optimal performance.

This effect is also illustrated by Lem. 1, where the performance loss of the asymptotic policy depends on $\ell_n(\pi_{\lambda^\star})$, which critically scales with the inverse of the spectral gap $\gamma_{\pi_{\lambda^\star}}$. This issue may also significantly affect the performance of FW-AME, as the gap between $\widehat{\psi}^+_{k+1}$ and $\widetilde{\psi}_{k+1}$ may be arbitrarily large if $\widehat{\pi}^+_{k+1}$ is slowly mixing. This problem together with Lem. 1 suggest regularizing the optimization problems (i.e., problem (5) for optimization and the computation of $\widehat{\psi}^+_{k+1}$ for learning) towards fast mixing policies.

Optimization. As a direct application of Lem. 1 we could replace problem (5) with

$$\min_{\substack{\pi \in \Pi^{\mathrm{SR}} \\ \eta \in \Delta(\mathcal{S})}} \mathcal{L}^{\mathrm{reg}}(\pi, \eta) := \mathcal{L}(\pi, \eta) + \ell_n(\pi) \quad \text{s.t. } \forall s \in \mathcal{S},\ \eta(s) = \sum_{s',a} \pi(a|s') p(s|s',a) \eta(s'). \tag{11}$$

The main advantage of solving this problem is illustrated in the following lemma.

Lemma 2. *Let $\pi^\star_{reg}$ be the solution of problem (11), its performance loss is bounded as*

$$\mathcal{L}_n(\pi^\star_{reg}) - \mathcal{L}_n(\pi^\star_n) \le 2\ell_n(\pi^\star_n). \tag{12}$$

Since in general we expect $\pi^\star_n$ to mix much faster than $\pi_{\lambda^\star}$ (i.e., $\gamma_{\pi^\star_n} \gg \gamma_{\pi_{\lambda^\star}}$), the performance loss of $\pi^\star_{\mathrm{reg}}$ may be much smaller than the loss in Lem. 1. As problem (11) is not convex, we replace it by heuristic convex algorithm. We isolate from $\ell_n(\pi)$ the spectral gap γ_π and the convergence rate $\rho_n := S/n$ and, using the norm formulation of the SLEM in (4), we introduce a proxy to the regularized loss as

$$\mathcal{L}(\pi, \eta) + \frac{\rho_n}{1 - \|D_\eta^{1/2} P_\pi D_\eta^{-1/2} - \sqrt{\eta}\sqrt{\eta}^{\mathsf{T}}\|_2}. \tag{13}$$

Building on this proxy and the study on computing fastest mixing chains on graphs by Boyd et al. (2004), we derive FMH (Faster-Mixing Heuristic) that solves a convex surrogate problem that favors fast mixing policies with limited deviation w.r.t. a target stationary distribution. While we postpone the full derivation to App. B.1, we report the main structure of the algorithm. FMH receives as input a budget n and the optimal asymptotic policy $\pi^\star$ obtained by solving (5), then it returns a stationary policy $\pi^\star_{\mathrm{FMH}}$. The algorithm proceeds through two steps.

Step 1 (improvement of the mixing properties). We first reparametrize the problem by introducing the variable $X \in \mathbb{R}^{S \times S}$ as $X = D_\eta P_\pi$ and we reduce the difficulty of handling the stationary constraint on η by constraining X to respect the adjacency matrix of the MDP Q. Notably, we introduce the constraints[5]

$$X = X^T, \quad X_{ss'} = 0 \ \text{ if } \ Q_{ss'} = 0, \tag{14}$$

which correspond to reversibility and adhering to the "structure" of the MDP. Furthermore, since we can recover a state distribution from X as $\eta_X(s) = \sum_{s'} X_{ss'}$, we also need to enforce

$$\sum_{s' \in \mathcal{S}} X_{ss'} \ge \underline{\eta}, \ \sum_{s \in \mathcal{S}} \Big(\sum_{s' \in \mathcal{S}} X_{ss'} - \eta^\star_s \Big)^2 \le \delta_n^2, \tag{15}$$

[5]We omit constraints $X \ge 0, \|X\|_1 = 1$ for clarity.

where we lower bound the state distribution and we require X to be close to the target state stationary distribution $\eta^\star$ in ℓ_2-norm. Since $\mathcal{L}$ is smooth when η is lower bounded by $\underline{\eta}$, the ℓ_2-norm constraint guarantees that the performance of X does not deviate much from $\eta^\star$. FMH then proceeds by solving

$$\min_X \sum_{s\in\mathcal{S}} \frac{\sigma^2(s)}{\sum_{s'\in\mathcal{S}} X_{ss'}} + \frac{\rho_n}{1 - \|D_{\eta^\star}^{-1/2} X D_{\eta^\star}^{-1/2} - \sqrt{\eta^\star}\sqrt{\eta^\star}^T\|_2}. \quad (16)$$
$$\text{s.t.} \quad (14), (15)$$

Unlike the proxy loss (13), this problem is convex in X and can be solved using standard convex optimization tools.

Step 2 (projection onto the set of feasible stationary policies). Unfortunately $\eta_X(s) = \sum_{s'} X_{ss'}$ may not be feasible in the MDP (i.e., it may not be stationary). Thus we finally proceed with the computation of a policy π whose stationary distribution is closest to η by solving the convex problem

$$\min_\pi \quad \sum_{s\in\mathcal{S}} \Big(\eta_X(s) - \sum_{s'\in\mathcal{S}, a\in\mathcal{A}_{s'}} \eta_X(s') p(s|s',a)\pi_{s',a}\Big)^2$$
$$\text{s.t.} \quad \pi_{s,a} \geq 0 \quad \text{and} \quad \sum_{a\in\mathcal{A}_s} \pi_{s,a} = 1.$$

FMH thus returns a policy that may have better mixing properties than $\pi^\star$ at the cost of a slight loss in asymptotic performance. The performance loss of FMH approaches the one of $\pi^\star_{\text{reg}}$ as shown in the following lemma.

Lemma 3. *Let $\pi^\star_{\text{FMH}}$ be the policy returned by* FMH*, then the performance loss is bounded as*

$$\mathcal{L}_n(\pi^\star_{\text{FMH}}) - \mathcal{L}_n(\pi^\star_n) \leq 2\ell_n(\pi^\star_n) + \frac{2\sigma^2_{\max}\sqrt{S}}{\underline{\eta}^2}\delta_n + \frac{2}{\gamma_{\min}}\rho_n + O(n^{-3/2}).$$

This suggests that the slack variable δ_n should decrease at least as $O(n^{-1})$ to guarantee the algorithm's consistency and not worsen the overall performance.

Finally, we introduce in App. B.2 a more computationally efficient variant of step 1 of FMH that uses semidefinite programming, which we later refer to as FMH-SDP.

Learning. As discussed above and shown in the proof of Thm. 1, the regret of FW-AME depends on the mixing properties of the policy $\widehat{\pi}^+_{k+1}$. While the optimization problem to compute $\widehat{\psi}^+_{k+1}$ is different than problem (5), the surrogate optimization procedure described above can be readily applied to this case as well. In fact, $\eta^\star$ received in input is now the target state-action stationary distribution $\widehat{\psi}^+_{k+1}$ and, since the objective function is still smooth, the deviation constraint does limit the performance loss that could be incurred because of the deviation δ_n. App. D.4 provides more discussion on the resulting learning algorithm that we call FW-AME w/ FMH-SDP.

5 Numerical Simulations

Experimental settings. We consider $\nu(s) = \mathcal{N}(0, \sigma^2(s))$ and when $T(s) = 0$, we set default variance and mean predictions to $\sigma^2_{\max}$ and $3\sigma_{\max}$. The initial state is drawn uniformly at random from $\mathcal{S}$. The episodes of FW-AME are set so that $t_k = \tau_1 + (k-1)^3$ (for $k > 1$, otherwise $t_1 = 1$), where τ_1 is the (adaptive) time needed for the initial policy to collect at least one sample of each state (so as to satisfy Asm. 1 after the first episode). We set $\underline{\eta} = 0.001$ and the confidence intervals to $\alpha(t, s, 1/t) = 0.2\sigma^2_{\max}\sqrt{\log(4St^2)/T_t(s)}$. We run simulations on a set of random Garnet MDPs (Bhatnagar et al., 2009). A Garnet instance $\mathcal{G}(S, A, b, \sigma^2_{\min}, \sigma^2_{\max})$ has S states, A actions, b is the branching factor and state variances are random in $[\sigma^2_{\min}, \sigma^2_{\max}]$. $\mathcal{G}_{\mathcal{R}}$ denotes the reversible Garnet MDPs (see App. E for more details). We set $\sigma^2_{\min} = 0.01$ and $\sigma^2_{\max} = 10$ to have a large spread between the state variances. For any budget n and policy π ran over R runs, the estimation loss is

$$\text{LOSS}(\pi, n, R) = \frac{1}{SR}\sum_{s\in\mathcal{S}}\sum_{1\leq r\leq R}\Big[\big(\widehat{\mu}^{(r)}_{\pi,n}(s) - \mu(s)\big)^2\Big],$$

while the normalized loss is $n\text{LOSS}(\pi, n, R)$. Finally, we measure the competitive ratio w.r.t. the optimal asymptotic performance as

$$\text{RATIO}(\pi, n, R) = \frac{n\text{LOSS}(\pi, n, R)}{\mathcal{L}(\lambda^\star)} - 1. \quad (17)$$

Results. We first verify the regret guarantees of Thm. 1. Fig. 3 reports the competitive ratio averaged over 100 randomly generated Garnet MDPs for FW-AME and a uniform policy $\pi_{\text{unif}}(a|s) = 1/|\mathcal{A}_s|$. As expected the ratio (which is a proxy for the regret) of FW-AME is much smaller than for π_{unif} and it approaches zero as the budget increases. While we report only the aggregated values, this result is consistently confirmed across all Garnet instances we have tried.

We then study the effectiveness of FMH in improving the optimization performance. In Fig. 2 we report $\text{LOSS}(\pi, n)$ for the asymptotic optimal policy $\pi_{\lambda^\star}$ and the surrogate policy $\pi^\star_{\text{FMH}}$ as a function of n in the simple 3-state MDP illustrated in Fig. 1, where $\pi^\star$ mixes poorly. We notice that in this case, the impact of favoring faster mixing policies does translate to a significant improvement in finite-time performance. This

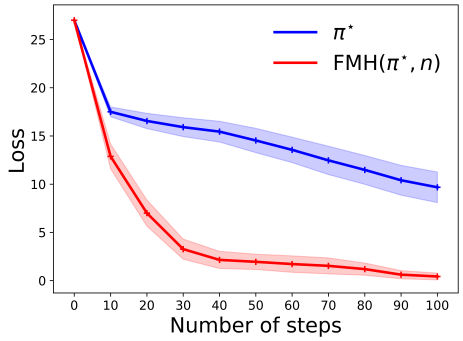
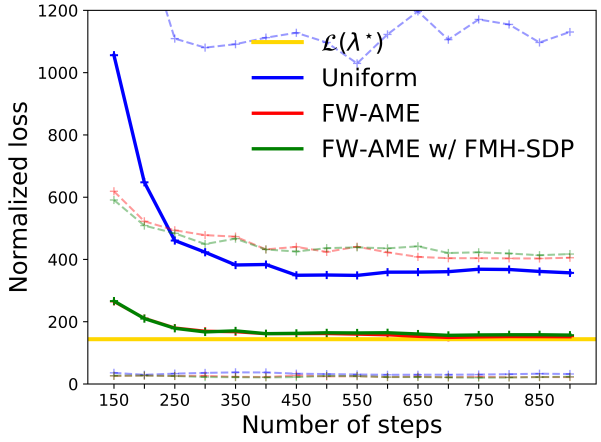


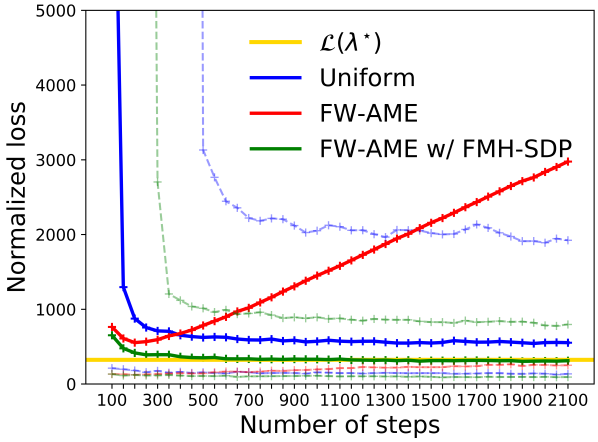
Figure 2: LOSS($\pi, n, R = 100$) as a function of n in the 3-state MDP of Fig. 1 (with $\sigma_2^2 = 0.001$).

π	$\pi_{\text{FW-AME}}$		π_{unif}	
n	500	1000	500	1000
$\mathcal{G}_{S=5}$	0.31	0.10	2.18	1.04
$\mathcal{G}_{S=10}$	0.35	0.19	1.98	1.15

Figure 3: RATIO($\pi, n, R = 100$) for $n \in \{500, 1000\}$ and for $\pi_{\text{FW-AME}}$ and π_{unif}, averaged over 100 Garnet instances randomly generated from $\mathcal{G}(S, A = 3, b = 2)$ for $S \in \{5, 10\}$.



(a) An instance of $\mathcal{G}_{\mathcal{R}}(S = 5, A = 3, b = 3)$ with fast mixing policies. The average SLEM is roughly 0.55, w/ or w/o FMH-SDP.



(b) An instance of $\mathcal{G}_{\mathcal{R}}(S = 10, A = 2, b = 2)$ where policies mix poorly. The average SLEM is 0.95 and it is decreased to 0.88 by FMH-SDP.

Figure 4: nLOSS($\pi, n, R = 1000$) as a function of n. The dashed curves report 5% and 95% quantiles.

finding is also confirmed when FMH is applied to FW-AME. We first show a specific reversible Garnet MDP where all the policies generated by FMH are mixing relatively fast (see the normalized loss in Fig. 4a). In this case, FMH-SDP has the same performance as FW-AME (and both are significantly better than uniform). This is confirmed by evaluating the average SLEM of the policies generated by the two algorithms, which is roughly 0.55 in both cases. On the other hand, there are Garnet MDP instances where FW-AME may indeed generate very poorly mixing policies that are executed for relatively long episodes, thus compromising the performance of the algorithm (see Fig. 4b).[6] In this case, FMH-SDP successfully biases the learning process towards faster mixing policies and obtains a much better finite-time performance. In fact, the average SLEM of the policies generated FW-AME is successfully reduced from 0.95 to 0.88 for FMH-SDP.

[6]The algorithm is still able to recover from bad mixing policies thanks to ergodicity and changing episodes, but it takes much longer to converge.

6 Conclusion and Extensions

We introduced the problem of active exploration in MDPs, proposed an algorithm with vanishing regret and proposed a heuristic convex optimization problem to favor fast mixing policies. This paper opens a number of questions: *(1)* A lower bound is needed to determine the complexity of active exploration in MDPs compared to the MAB case; *(2)* While the ergodicity assumption is not needed in regret minimization in MDPs (Jaksch et al., 2010), it is unclear whether it is mandatory in our setting; *(3)* A full regret analysis of the case of unknown MDP (see App. A.3). This paper may be a first step towards formalizing the problem of intrinsically motivated RL, where the implicit objective is often to accurately estimate the MDP dynamics and effectively navigate through states (see e.g., Auer et al., 2011; Hazan et al., 2018).

References

Akshay, S., Bertrand, N., Haddad, S., and Helouet, L. (2013). The steady-state control problem for markov decision processes. In *International Conference on Quantitative Evaluation of Systems*, pages 290–304.

Antos, A., Grover, V., and Szepesvári, C. (2010). Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728.

Auer, P., Lim, S. H., and Watkins, C. (2011). Models for autonomously motivated exploration in reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 14–17.

Balcan, M.-F., Blum, A., Haghtalab, N., and Procaccia, A. D. (2015). Commitment without regrets: Online learning in stackelberg security games. In *Proceedings of the sixteenth ACM conference on economics and computation*, pages 61–78.

Basilico, N., Gatti, N., and Amigoni, F. (2012). Patrolling security games: Definition and algorithms for solving large instances with single patroller and single intruder. *Artificial Intelligence*, 184:78–123.

Berthet, Q. and Perchet, V. (2017). Fast rates for bandit optimization with upper-confidence frank-wolfe. In *Advances in Neural Information Processing Systems*, pages 2225–2234.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.

Boyd, S., Diaconis, P., and Xiao, L. (2004). Fastest mixing markov chain on a graph. *SIAM review*, 46(4):667–689.

Carpentier, A., Lazaric, A., Ghavamzadeh, M., Munos, R., and Auer, P. (2011). Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 189–203.

Chentanez, N., Barto, A. G., and Singh, S. P. (2005). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17*, pages 1281–1288.

Dance, C. R. and Silander, T. (2017). Optimal policies for observing time series and related restless bandit problems. *arXiv preprint arXiv:1703.10010*.

Diaconis, P., Stroock, D., et al. (1991). Geometric bounds for eigenvalues of markov chains. *The Annals of Applied Probability*, 1(1):36–61.

Hanneke, S. (2014). Theory of disagreement-based active learning. *Foundations and Trends in Machine Learning*, 7(2-3):131–309.

Hazan, E., Kakade, S. M., Singh, K., and Soest, A. V. (2018). Provably efficient maximum entropy exploration. *CoRR*, abs/1812.02690.

Hsu, D. J., Kontorovich, A., and Szepesvári, C. (2015). Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *Proceedings of The 30th International Conference on Machine Learning*, volume 28, pages 427–435.

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

Lattimore, T. and Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge University Press (preprint).

Paulin, D. et al. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20.

Pukelsheim, F. (2006). *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA.

Rolf, E., Fridovich-Keil, D., Simchowitz, M., Recht, B., and Tomlin, C. (2018). A successive-elimination approach to adaptive robotic sensing. *CoRR*, abs/1809.10611.

Welch, W. J. (1982). Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15(1):17–25.