Bayesian Optimistic Optimization: Optimistic Exploration for Model-based Reinforcement Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Reinforcement learning (RL) is a general framework for modeling sequential decision making problems, at the core of which lies the dilemma of exploitation and exploration. An agent failing to explore systematically will inevitably fail to learn efficiently. Optimism in the face of uncertainty (OFU) is a conventionally successful strategy for efficient exploration. An agent following the OFU principle explores actively and efficiently. However, when applied to model-based RL, it involves specifying a confidence set of the underlying model and solving a series of nonlinear constrained optimization, which can be computationally intractable. This paper proposes an algorithm, Bayesian optimistic optimization (BOO), which adopts a dynamic weighting technique for enforcing the constraint rather than explicitly solving a constrained optimization problem. BOO is a general algorithm shown to be sample-efficient for various model classes. We also developed effective optimization techniques based on natural gradients and entropy regularization.

1 Introduction

Reinforcement learning (RL) is a sequential decision-making problem in which an agent acts in an unknown environment while maximizing the cumulative rewards it receives. In this paper, we consider the RL in Markov decision processes (MDPs), where the agent observes the state of the environment at each timestep and makes decisions accordingly. Since the environment is unknown, maximizing the cumulative rewards naturally involves a trade-off between exploration and exploitation. Exploitation is to make the best-rewarding decision based on the information the agent possesses, and exploration means actively gathering information about the environment so that the agent understands better about the environment and thus makes better decisions in the future. An algorithm cannot be sample-efficient without balancing them properly.

Theoretically, the exploration and exploitation dilemma admits a Bayesian optimal solution [1]. That is to consider the RL problem a so-called Bayes-Adaptive MDP (BAMDP), a special case of partially observable Markov decision processes (POMDPs), where the parameter of the dynamics is unobservable. Although this formulation provides useful insights, the POMDP formulation is computationally intractable [2, 3]. Therefore, all practical algorithms resolve this dilemma by achieving a delicate balance between seeking reward and gathering information. Optimism in the face of uncertainty (OFU) is one of the conventionally successful approaches for this balance and is established as an efficient learning principle in various cases [4–6].

The OFU principle makes the agent an optimist who is always optimistic about the uncertainty of the environment. If it ever stands a chance that a policy is highly profitable, the agent will try it out. By executing an optimistic policy, the agent either receives high rewards or gains information after observing unexpected outcomes. OFU involves constructing a confidence set of possible MDPs and solving for the most optimistic one within the confidence set. Unfortunately, this is typically

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

a constrained nonlinear non-concave optimization and is impractical to solve [7, 8]. Therefore, a model-based OFU algorithm is previously deemed an inferior choice [8].

In this paper, we propose a model-based OFU algorithm, Bayesian optimistic optimization (BOO), which enforces the constraint via a dynamic weighting technique and therefore makes the optimization practical. We show that BOO is a general-purpose model-based OFU RL algorithm that is both provably sample-efficient and comparatively computationally tractable. Our contribution is three-fold:

- We extend the OFU algorithms to the Bayesian setting by requiring the model to reside in a Bayesian credible region. BOO is then proposed for enforcing such requirement.
- We prove that BOO is efficient in terms of the frequentist regret for various model classes.
- We derive an optimization method for BOO based on the natural gradient and entropy regularization.

2 Preliminaries

This section introduces the preliminaries, including RL in MDPs and the principle of optimism in the face of uncertainty.

2.1 Reinforcement Learning in Markov Decision Processes

We consider the problem where a reward-seeking agent repeatedly interacts with a finite-horizon MDPs $M = (\mathcal{S}, \mathcal{A}, P, R, H, s_1)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively. The algorithm adopted by the agent is denoted as $\mathfrak{A}$. At each episode, the agent is spawned at the initial state $s_1 \in \mathcal{S}$. It takes an action $a_h \in \mathcal{A}$ at each period h within an episode and receives a random reward, $r_h \sim R(s_h, a_h)$. Then, the environmental state transitions to the next state, $s_{h+1} \sim P(s_h, a_h)$. This process repeats until the episode ends at the H-th period, and another episode begins. Note that the fixed initial state is not a stringent condition since starting from an initial distribution ρ_0 is equivalent to starting from a special state s_1 such that $P(s|s_1, a) = \rho_0(s)$ holds for any state s and action a. A finite MDP is an MDP with finite state and action spaces.

A policy π is a function mapping a state and a period to an action distribution. The value function for the MDP M, policy π, and period h is defined recursively as

$$\begin{aligned} V_h^{\pi,M}(s_h) &= \mathbb{E}_{a_h \sim \pi_h(s_h)} \left[Q_h^{\pi,M}(s_h, a_h) \right], \forall h \in [H], \\ Q_h^{\pi,M}(s_h, a_h) &= \bar{R}^M(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim P^M(s_h, a_h)} \left[V_{h+1}^{\pi,M}(s_{h+1}) \right], \forall h \in [H-1], \end{aligned} \tag{1}$$

where $\bar{R}^M(s, a) = \mathbb{E}_{r \sim R^M(s,a)}[r]$, and $Q_H^{\pi,M}(s_H, a_H) = \bar{R}^M(s_H, a_H)$.

We denote by $t = (k-1)H + h$ the number of the total time steps at the h-th period of the k-th episode. The history prior to t is defined as $\mathcal{H}_t = (s_1, a_1, r_1, s_2, \ldots, s_{t-1}, a_{t-1}, r_{t-1})$. For notational convenience, we also use X_{kh} to represent X_t, where $t = (k-1)H + h$. The agent is assumed to be capable of memorizing the entire history. It produces a policy π_k and a model M_k per episode in light of the history $\mathcal{H}_{k1}$, where $k1$ stands for $(k-1)H + 1$. In this paper, we only consider cases where the policy π_k is an optimal policy of the MDP M^k, i.e., $\pi_k \in \arg\max_\pi V_1^{\pi,M_k}(s_1)$. For any history $\mathcal{H}$, $\mathfrak{A}(\mathcal{H})$ defines a distribution over models and policies. Within the k-th episode, the agent samples an action from $\pi_k(s_{kh}, h)$ at each period h.

In the frequentist viewpoint, there exists an unknown true MDP M^*, and the performance metric, regret, is defined in terms of this true MDP:

$$\mathrm{Regret}(T, \mathfrak{A}, M^*) = \mathbb{E}_{\mathcal{H}_T \sim \mathfrak{A}, M^*} \left[\sum_{k=1}^{K} \mathbb{E}_{\pi_k \sim \mathfrak{A}(\mathcal{H}_{k1})} \left[V_1^*(s_1) - V_1^{\pi_k, M^*}(s_1) \right] \right] = \sum_{k=1}^{K} \Delta_k, \tag{2}$$

where $V_h^* = V_h^{\pi^*, M^*}$, π^* is an optimal policy of the true MDP M^*, K is the total number of episodes, $T = KH$ is the number of the total time steps, and $\Delta_k = \mathbb{E}[V_1^*(s_1) - V_1^{\pi_k, M^*}(s_1)]$. In the Bayesian viewpoint, the unknown MDP is treated as a random variable and assigned a prior ρ_M. All MDPs in the support of ρ_M differ only in the transition function P and the reward function R. The Bayesian objective of the agent is to minimize the Bayesian regret up to time T, $\mathrm{BayesRegret}(T, \mathfrak{A}, \rho_M) = \mathbb{E}_{M^* \sim \rho_M} [\mathrm{Regret}(T, \mathfrak{A}, M^*)]$.

2.2 Optimism in the Face of Uncertainty

Optimism in the face of uncertainty is a strategy for information gathering. When the optimal action is not clear given the current information, it is preferable to hazard an optimistic guess. If we make an atrocious guess, we effectively rule that out and pick another next time. Otherwise, we end up finding a competitive solution that incurs little regrets. This idea is mathematically realized as a constrained optimistic optimization, $\max_{\pi_k, M_k} V_1^k(s_1) \quad \text{s.t. } M_k \in \mathcal{M}_k$, where $\mathcal{M}_k$ is a confidence set constructed using empirical data $\mathcal{H}_{k1}$ such that $M^* \sin \mathcal{M}_k$ with high probability. The pseudocode of the OFU algorithm is shown in Algorithm 1.

Algorithm 1 OFU RL

1: **for** episode $k = 1, 2, \ldots$ **do**
2: Construct a confidence set $\mathcal{M}_k$ with $\mathcal{H}_{k1}$
3: Compute $\pi_k \in \arg\max_\pi \max_{M_k} V_1^{\pi, M_k}(s_1) \quad \text{s.t.} \quad M_k \in \mathcal{M}_k$
4: Execute π_k for an episode

3 Related Work

Our algorithm is designed following a Bayesian perspective. It extends frequentist OFU algorithms to Bayesian settings by using a Bayesian credible region. This idea is first proposed by [9] in bandit problems. Their method replaces frequentist reward confidence interval with posterior reward quantile and is provably efficient in terms of the frequentist measure of performance. [10] extends their approach to RL by modeling the transition distribution as set-valued Bernoulli distribution. We further generalize this idea to use any Bayesian credible region in place of the frequentist confidence set.

BOO balances exploration and exploitation via optimizing both the value and the log-posterior density. We note that, when the prior is uniform, this idea is equivalent to balance value versus log-likelihood, which is first explored in [11] and is named as reward-biased maximum likelihood estimation (RBMLE). They have applied this approach to multi-armed bandits [12], contextual bandits [13], and RL where the model belongs to a known finite set [11, 14]. A constrained version of RBMLE is also successfully applied to RL of linear quadratic control systems [15]. Our algorithm can be considered generalizing RBMLE to a Bayesian perspective and general model class. In this regard, BOO could also be referred to as reward-biased maximum a posteriori (RBMAP).

4 Bayesian Optimistic Optimization

In this section, we derive the learning objective of BOO as a Lagrangian relaxation of a constrained optimization problem and give an intuitive interpretation of the resulting objective.

4.1 Constrained Bayesian Optimistic Optimization

The conventional OFU algorithm, as demonstrated in Algorithm 1, contains a constrained optimistic optimization, where we look for an optimistic MDP $M_k \in \mathcal{M}_k$ and its corresponding optimal policy such that the value is maximized. The constrained Bayesian optimistic optimization (CBOO) is almost the same (see Algorithm 2), except the frequentist confidence set is now replaced with the Bayesian credible region. A credible region with a $1 - \alpha_k$ level of confidence is a set $\mathcal{M}_k$ such that $\Pr(\mathcal{M}_k | \mathcal{H}_{k1}) \geq 1 - \alpha_k$, where $\Pr(\cdot | \mathcal{H}_{k1})$ is the posterior distribution given history $\mathcal{H}_{k1}$, and $\Pr(\mathcal{M}_k | \mathcal{H}_{k1}) = \int_{M_k \in \mathcal{M}_k} \Pr(M_k | \mathcal{H}_{k1}) \, \mathrm{d}M_k$.

By the construction of the credible region, we have $M^* \in \mathcal{M}_k$ holds with probability $1 - \alpha_k$ given any history $\mathcal{H}_{k1}$. Therefore, the per-episode Bayesian regret is bounded with probability $1 - \alpha_k$,

$$\mathbb{E}[\Delta_k] \leq \underbrace{\mathbb{E}\left[V_1^*(s_1) - V_1^k(s_1) \middle| M^* \in \mathcal{M}_k\right]}_{\tilde{\Delta}_k^{\text{opt}}} + \underbrace{\mathbb{E}\left[V_1^k(s_1) - V_1^{\pi_k, M^*}(s_1) \middle| M^* \in \mathcal{M}_k\right]}_{\tilde{\Delta}_k^{\text{conc}}} \leq \tilde{\Delta}_k^{\text{conc}}, \tag{3}$$

where $V_h^k = V_h^{\pi_k, M_k}$, and the optimism term $\tilde{\Delta}_k^{\text{opt}}$ is less than or equal to 0 by construction. We define a distance metric $d(M_1, M_2) = \max_\pi |V_1^{\pi,M_1}(s_1) - V_1^{\pi,M_2}(s_1)|$. It is preferable to have a credible region $\mathcal{M}$ such that the set width $\max_{M_1,M_2\in\mathcal{M}} d(M_1, M_2)$ is minimized since the set width certifies an upper bound on $\tilde{\Delta}_k^{\text{conc}}$. However, designing a value concentration credible region could be intractable for generic model classes. As a reasonable alternative, we propose the highest density region (HDR) [16] or, in the Bayesian context, the highest posterior density (HPD) region [17], i.e., $\mathcal{M}_k = \{M_k | \Pr(M_k|\mathcal{H}_{k1}) \geq \epsilon_k\}$, where ϵ_k is the largest constant such that $\Pr(\mathcal{M}_k|\mathcal{H}_{k1}) \geq 1 - \alpha_k$. This kind of region features a desirable property that it occupies the smallest volume in the sample space among all credible regions of the same confidence level and has a potentially small set width.

There are two problems preventing the CBOO from being a practical algorithm. The first is that the ϵ_k in the definition of the HPD region is unknown. Although we may approximate it with the Monte Carlo approximation [18], an estimator of high/low quantiles has high variance rendering the approximation difficult. The other difficulty is that the constrained joint optimization of model and policy is an NP-hard problem even in bandits with linear reward and quadratic constraints [7].

Algorithm 2 CBOO

for episode $k = 1, 2, \ldots$ **do**
  Construct a **credible region** $\mathcal{M}_k$ with $\mathcal{H}_{k1}$
  Compute $\pi_k \in \arg\max_\pi \max_{M_k} V_1^{\pi,M_k}(s_1) \quad \text{s.t.} \quad M_k \in \mathcal{M}_k$
  Execute π_k for an episode

4.2 Lagrangian Relaxation of CBOO

By introducing a Lagrange multiplier λ_k, we transform the CBOO with HPD regions into an unconstrained optimization problem, $\max_{\pi,M} \left(V_1^{\pi,M}(s_1) + \lambda_k(\log \Pr(M|\mathcal{H}_{k1}) - \log \epsilon_k) \right)$. Once the Lagrange multiplier is determined, the constant $\log \epsilon_k$ and the marginal likelihood $\log \Pr(\mathcal{H}_{k1})$ are irrelevant, and the optimization is equivalent to $\max_{\pi,M} \left(V_1^{\pi,M}(s_1) + \lambda_k(\log \Pr(\mathcal{H}_{k1}|M) + \log \Pr(M)) \right)$. This gives rise to the BOO algorithm as shown in Algorithm 3.

Algorithm 3 Bayesian Optimistic Optimization

for episode $k = 1, 2, \ldots$ **do**
  Compute $\pi_k \in \arg\max_\pi \max_M \left(V_1^{\pi,M}(s_1) + \lambda_k(\log \Pr(\mathcal{H}_{k1}|M) + \log \Pr(M)) \right)$
  Execute π_k for an episode

However, the problem is how to determine the value of λ_k. We note that the following argument by posterior sampling provides an inexact yet insightful viewpoint. A strict discussion is presented in Section 5. The maximization of the BOO objective can be considered the problem of selecting the best one among multiple posterior samples (see Algorithm 4). When only one sample is taken, Algorithm 4 reduces to the well-known posterior sampling for reinforcement learning (PSRL) [19]. As the number of samples d goes to infinity, maximizing among all posterior samples gives approximately the solution of the above optimization.

Algorithm 4 BOO via Posterior Sampling

for episode $k = 1, 2, \ldots$ **do**
  Sample $M_k^1, M_k^2, \ldots, M_k^d \sim \Pr(\cdot|\mathcal{H}_{k1})$
  Compute $\pi_k \in \arg\max_\pi \max_{i\in[d]} \left(V_1^{\pi,M_k^i}(s_1) + \lambda_k(\log \Pr(\mathcal{H}_{k1}|M_k^i) + \log \Pr(M_k^i)) \right)$
  Execute π_k for an episode

Notice that we can regard Algorithm 4 as maximizing over several identically distributed random variables, $\lambda_k \log \Pr(M_k^i|\mathcal{H}_{k1}) + V_1^{\pi_k^i, M_k^i}(s_1)$. The magnitude of their mean does not matter since

subtracting a constant will not change the optimum. The thing that matters is their variation. We need to make sure that the variation is not dominated by either $V_1^{\pi_k^i, M_k^i}(s_1)$ or $\lambda_k \log \Pr(M_k^i|\mathcal{H}_{k1})$. If the value dominates the variation, the algorithm will select an unreliable model, which causes inefficiency. If dominated by the probability, it shows too strong a preference for high probability models and hesitates to explore.

Theorem 4.1 shows that the standard deviation of log probability is at least a constant (proof in Appendix B).

Theorem 4.1 (Variation of the log-posterior density)**.** *Suppose $X^n = (X_1, X_2, \ldots, X_n)$ are observations from a stochastic process whose distribution P_θ depends on $\theta \in \Theta$, an open subset of $\mathbb{R}^m$. Assume that the posterior is asymptotic normal, and the log-posterior density is continuous. The variance of the log-posterior density satisfies* $\liminf_{n\to\infty} \mathrm{Var}_\theta \left[\log \Pr(\theta|X^n)\right] \geq m/2$.

The process where an algorithm $\mathfrak{A}$ interacts with a random MDP parameterized by θ is a stochastic process where the observation X_t is the state-action pair (s_t, a_t), and the observation distribution is $\Pr(X^n|\theta, \mathfrak{A})$ is a distribution depending on θ. Thus, by Theorem 4.1, we know that if the posterior distribution is asymptotic normal, then, for a random sample θ from the posterior distribution, the variance of the random variable $\log \Pr(\theta|X^n)$ is at least $m/2$ when n is large enough.

Theorem 4.1 relies on the asymptotic normality of the posterior distribution, which is satisfied under some regularity conditions specified in [20]. It is also possible to strengthen this result and derive a convergence of the variance of log-posterior density. Indeed, it is empirically observed that this quantity converges rapidly to a constant.

For a proper RL algorithm, it is clear that the standard deviation of $V_1^{\pi_k^i, M_k^i}(s_1)$ should shrink w.r.t. k. Otherwise, the algorithm fails to find out the optimal value and policy. Suppose that the standard deviation of $V_1^{\pi_k^i, M_k^i}$ is $O(k^{-1/\alpha})$. We need to set the scaling parameter λ_k proportionally such that the variation of the log-posterior density term $\log \Pr(M_k^i|\mathcal{H}_{k1})$ matches that of the value term $V^{\pi_k^i, M_k^i}$. Indeed, given a specific model class and noise type, it is possible to derive a worst-case rate for the shrinkage of the value uncertainty and thus determine the proper value of λ_k. Nonetheless, the rate of information revealing could appear in an instance-dependent manner. That is to say, the optimal policy of some MDPs could be inherently easier to determine than the others. Even in a fixed MDP, the rate of uncertainty shrinkage could also change abruptly. For example, consider a ReLU bandit problem, where $B = \{p \in \mathbb{R}^d | \|p\|_2 \leq 1\}$, and the agent at each episode selects an action $a \in B$ and receives a Gaussian reward of mean $\max(\theta^\top a, 0)$. If it happens that the agent selects an action a in the inactive region where $\theta^\top a \leq 0$, the resulting observation reveals little information about the optimal action. Suppose the parameter θ is selected such that the positive-reward action region is of a maximum width $\epsilon > 0$. Then, in the worst case, the agent needs to explore $\Omega(1/\epsilon^{d-1})$ actions in order to find the positive reward region. However, the uncertainty about the optimal action starts diminishing rapidly whenever the positive reward region is identified.

The view of matching the log probability variation with the value uncertainty provides rationales for the dynamic adjustment of λ_k. Meanwhile, it points out the potential limitation of decaying λ_k with a fixed rate. In this paper, we focus on methods that scale λ_k with a fixed rate, but as mentioned above, it would be fascinating to adjust it in an instance-dependent manner.

5 Regret Analysis

This section analyzes the regret of BOO for general model classes. To begin with, we introduce complexity measurement for function classes.

5.1 Complexity Measurement

Our regret analysis uses the notions of the covering number and the eluder dimension [21, 22]. They measure different aspects of function classes.

The covering number is introduced for measuring the size of the function class.

Definition 1 (Covering number)**.** Let $(M, \|\cdot\|)$ be a metric space and $\mathcal{F} \subseteq M$. The covering number $N(\mathcal{F}, \alpha, \|\cdot\|)$ for $\alpha > 0$ is the minimum cardinality of $C \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$, there exists $c \in C$ satisfying $\|f - c\| \leq \alpha$.

The eluder dimension is a complexity measure for the exploration difficulty. Similar to [23], we generalize the original eluder dimension [21] to the distributional eluder dimension by measuring the ℓ_2 difference on a distribution.

Definition 2 (ϵ-dependence between distributions [22, 23])**.** Let $\mathcal{F}$ be a function class defined on $\mathcal{X}$, and $\nu, \mu_1, \ldots, \mu_n$ be probability measures over $\mathcal{X}$. We say ν is ϵ-dependent on $\{\mu_1, \mu_2, \ldots, \mu_n\}$ w.r.t. $\mathcal{F}$ if any $f_1, f_2 \in \mathcal{F}$ satisfying $\sqrt{\sum_{i=1}^n \mu_i(\|f_1 - f_2\|_2)^2} \leq \epsilon$ also satisfies $\nu(\|f_1 - f_2\|_2) \leq \epsilon$. Here, $\mu(f) = \int f \,\mathrm{d}\mu$. If ν is not ϵ-dependent on $\{\mu_1, \mu_2, \ldots, \mu_n\}$, then it is said to be ϵ-independent of $\{\mu_1, \mu_2, \ldots, \mu_n\}$ w.r.t. $\mathcal{F}$.

Let f^* be the real function where observations are generated. Intuitively, a distribution ν is ϵ-independent of $\{\mu_1, \mu_2, \ldots, \mu_n\}$ means that a function f indistinguishable from the real function f^* on historical distributions could still be significantly different on the distribution ν.

Definition 3 (Distributional eluder (DE) dimension [22, 23])**.** Let $\mathcal{F}$ be a function class defined on $\mathcal{X}$, and Π be a class of probability measures over $\mathcal{X}$. The distributional eluder dimension $\mathrm{DEdim}(\mathcal{F}, \Pi, \epsilon)$ is the length of the longest sequence in Π such that every element in the sequence is ϵ'-independent of its predecessors for some $\epsilon' \geq \epsilon$.

The distributional eluder dimension states that the bad event that a historically indistinguishable function f fails to match the real function on future data cannot happen indefinitely. In the worst case, after $\mathrm{DEdim}(\mathcal{F}, \Pi, \epsilon)$ such events, all historically indistinguishable function matches the real function with ϵ-precision. The distributional eluder dimension equals the original eluder dimension when Π is the set of Dirac delta distribution centered around each point in the function's domain. In RL, the distribution set we are concerned with is the distributions induced by the models.

Definition 4 (Probability measures induced by models)**.** Let $\mathcal{M}$ be a set of models and M^* be the real MDP. The probability measures induced by $\mathcal{M}$ and M^* is $\Pi_{\mathcal{M}}^{M^*} = \{\rho^{\pi,M} = \frac{1}{H}\sum_{h=1}^H \rho_h^{\pi,M} | M \in \mathcal{M}, \pi \in \arg\max V_1^{\pi,M}(s_1)\}$, where $\rho_h^{\pi,M}$ denotes the state-action distribution at period h induced by executing an optimal policy π of M in the MDP M^*.

We consider the following function classes:

1. Linear models: $\mathcal{F} = \{f_\theta | \theta \in \Theta, f_\theta(x) = \theta\phi(x), \forall x \in \mathcal{X}\}$, where $\Theta \subset \mathbb{R}^{m\times d}$ and $\|\theta\|_2 \leq r_\theta$ for all $\theta \in \Theta$, $\phi(x) \in \mathbb{R}^d$, and $\|\phi(x)\|_2 \leq r_\phi$ for all $x \in \mathcal{X}$.
2. Generalized linear model: $\mathcal{F} = \{f_\theta | \theta \in \Theta, f_\theta(x) = g(\theta\phi(x)), \forall x \in \mathcal{X}\}$, where $\Theta \subset \mathbb{R}^{m\times d}$, $\|\theta\|_2 \leq r_\theta$ for all $\theta \in \Theta$, $\phi(x) \in \mathbb{R}^d$, and $\|\phi(x)\|_2 \leq r_\phi$ for all $x \in \mathcal{X}$. Assume there exist some $\overline{h}$ and $\underline{h}$ such that the gradient of g satisfies $0 < \underline{h} \leq g'(x) \leq \overline{h}$ for all $x \in \mathbb{R}^d$ satisfying $\|x\|_2 \leq r_\theta r_\phi$.
3. Reproducing kernel Hilbert space (RKHS) ball: $\mathcal{F} = \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq r_{\mathcal{H}}\}$, where $\mathcal{H}$ is vector-valued RKHS on a Lebesgue space $(\mathcal{X}, \Sigma, \mu)$ with a separable kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{m\times m}$ constructed by $K(x, y)_{ij} = K_{\mathcal{X}}(x, y)K_m(i, j)$. Here, $K_{\mathcal{X}}$ and K_m are kernels for scalar-valued RKHS on $\mathcal{X}$ and $[m]$, respectively. The RKHS norm $\|\cdot\|_{\mathcal{H}}$ is $\|f\|_{\mathcal{H}} = (\int \|f(x)\|_2^2 \,\mathrm{d}\mu)^{1/2}$.

Table 1 lists the complexity of different function classes measured by eluder dimension and covering number. Here, we extend previous covering number bound for scalar-valued functions [22, 24] to vector-valued functions. The upper bounds for distributional eluder dimension is first established. The complexity of RKHS relies on the eigenvalue decay property of the kernel. The finite dimensional linear kernel, the squared exponential kernel, and the Martérn kernel are examples of the finite spectrum, exponential decay, and polynomial decay kernels, respectively. The detailed explanation and derivation are in Appendix A.

5.2 BOO Regret

This section analyze the regret of BOO. Instead of general MDPs, we now work with a specific type of MDPs consisting of a model function and an additive sub-Gaussian noise.

Table 1: Log covering number and distributional eluder dimension

	$\log N(\mathcal{F}, \alpha, \|\cdot\|_\infty)$	$\mathrm{DEdim}(\mathcal{F}, \Pi, \epsilon)$
Linear Model	$O(md\log(r_\phi r_\theta/\alpha))$	$O(md\log(r_\phi^2 r_\theta^2/\epsilon^2))$
Generalized Linear Model	$O(md\log(\overline{h} r_\phi r_\theta/\alpha))$	$O(md(\overline{h}/\underline{h})^2\log(r_\phi^2 r_\theta^2 \overline{h}^2/\epsilon^2))$
γ-finite spectrum RKHS	$O(m\gamma\log(r_{\mathcal{H}}/\alpha))$	$O(m\gamma\log(1/\epsilon))$
γ-exponential decay RKHS	$O(m(\log(r_{\mathcal{H}}/\alpha))^{1+1/\gamma})$	$O((m\log(1/\epsilon))^{1+1/\gamma})$
γ-polynomial decay RKHS	$O(m(r_{\mathcal{H}}/\alpha)^{2/(\gamma(1-2\sigma)-1)}\log(r_{\mathcal{H}}/\alpha))$	$O(m(1/\epsilon)^{(d+1)/(\gamma+d)}\log(1/\epsilon))$

Assumption 1 (Models with additive sub-Gaussian noises)**.** *Let the state space $\mathcal{S}$ be a subset of $\mathbb{R}^d$. Denote the state-action space $\mathcal{S}\otimes\mathcal{A}$ as $\mathcal{X}$ and state-reward space $\mathcal{S}\otimes\mathbb{R}$ as $\mathcal{Y}$. Let $\mathcal{F}\subset\mathcal{Y}^{\mathcal{X}}$ be a class of model functions. For each $y\in\mathcal{Y}$, ϵ_y is a σ-sub-Gaussian noise ϵ_y, i.e., $\log\mathbb{E}[e^{v^\top\epsilon_y}]\le\|v\|_2^2\sigma^2/2$ for all $v\in\mathbb{R}^d$. There exists a function $f^*:\mathcal{X}\to\mathcal{Y}$ in $\mathcal{F}$ such that for any $x\in\mathcal{X}$, $f^*(x)+\epsilon_{f^*(x)}$ is distributed according to the joint distribution of $P^{M^*}(x)$ and $R^{M^*}(x)$.*

Assumption 1 requires the output distribution to be determined by its mean. This requirement is satisfied when the output distribution is modeled as, for example, Gaussian with known variance. The categorical distribution also satisfies this requirement by representing each category as a one-hot vector, in which case the mean is just the probability vector specifying the entire distribution.

Since the noise is assumed sub-Gaussian, we adopt the independent Gaussian likelihood as a surrogate for the actual likelihood function [13], i.e.,

$$\log\Pr(y|x,f)=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\|y-f(x)\|_2^2. \tag{4}$$

In this way, we avoid restricting our discussion on specific types of distributions.

The second assumption is the Lipschitz continuity of the one-step value function, which is needed for connecting the model difference to the value difference. Roughly speaking, this requires that when the outputs of model functions are close, the corresponding transition distributions are close.

Assumption 2 (One-step Value Lipschitz Continuity)**.** *Let $U_{x,V}(M)$ be the one-step value function with the state-action pair $x\in\mathcal{X}$ and the value function $V:\mathcal{S}\to\mathbb{R}$, i.e., $U_{x,V}(M)=\bar{R}^M(x)+P_x^M(V)$, where $P_x(V)=\int V(s')P(s'|x)\,\mathrm{d}s'$. We define the Lipschitz constant w.r.t. the 2-norm as*

$$|U_{x,V}(M)-U_{x,V}(M')|\le L_{x,V}\|f^M(x)-f^{M'}(x)\|_2, \tag{5}$$

where f^M is the model function of M. Let $\mathcal{M}$ be a set of MDPs. The global Lipschitz constant w.r.t. $\mathcal{M}$ is $L=\max_{x,\pi,M\in\mathcal{M}}L_{x,V^{\pi,M}}$.

Finally, we derive the asymptotic regret of the BOO algorithm in Theorem 5.1. Notice that the regret of BOO relies on both the complexity of model classes measured by the distributional eluder dimension and the log covering number and the decay rate of the scaling parameter, which conforms to our previous discussion.

Theorem 5.1 (Bayesian optimistic optimization regret)**.** *Assume that $\|f(x)\|_2\le r_{\mathcal{F}}$ for all $x\in\mathcal{X}$ and $f\in\mathcal{F}$, and the log-prior probability uniformly bounded by Z. The distributional eluder dimension is assumed to satisfy $\mathrm{DEdim}(\mathcal{F},\Pi_{\mathcal{M}}^{M^*},\epsilon)\le E(1/\epsilon)^{e_1}(\log(1/\epsilon))^{e_2}$ for $e_1,e_2\ge0$, and the log covering number satisfies $\log N(\mathcal{F},\alpha,\|\cdot\|_\infty)\le C(1/\alpha)^{c_1}(\log(1/\alpha))^{c_2}$ for $c_1,c_2\ge0$. Then, setting $\lambda_k=ck^{-v}$ for $c>0$, $e>0$, the frequentist regret of BOO is*

$$\sum_{k=1}^K\Delta_k=O\left(Hr_{\mathcal{F}}^2CK^{\frac{c_1}{c_1+2}+1-v}(\log K)^{2c_2/(c_1+2)}\right), \tag{6}$$

if $v<\frac{(e_1+1)c_1}{(e_1+2)(c_1+2)}+\frac{1}{e_1+2}$, and

$$\sum_{k=1}^K\Delta_k=O\left(L\left(H^2r_{\mathcal{F}}^2C\right)^{\frac{1}{e_1+2}}K^{1+\frac{v-1}{e_1+1}}(\log K)^{\frac{e_2}{e_1+1}}\right) \tag{7}$$

if otherwise.

As suggested by Theorem 5.1, the best performance of BOO is achieved by setting the decay rate to $v = \frac{(e_1+1)c_1}{(e_1+2)(c_1+2)} + \frac{1}{e_1+2}$.

6 Optimization

This section discusses the optimization of the BOO algorithm.

Algorithm 4 provides a method of optimizing the BOO objective via posterior sampling. For infinite-dimensional RKHS, this is the only choice. Since PSRL, the single-sample version of Algorithm 4, already achieves the state-of-the-art performance, combining multiple samples may seem meaningless. We argue that the potential advantage of Algorithm 4 is that, unlike PSRL, it does not require posterior samples to be independent. Even the requirement that samples are from the posterior can be relaxed. These features are essential either when the posterior distribution is approximated or when the posterior samples are produced by Markov chain Monte Carlo methods and are correlated.

For finite-dimensional models, we can optimize via gradient-based methods. The objective function of BOO consists of two parts. The gradient of the log-posterior density part w.r.t. the model is just the gradient of the log-prior and the log-likelihood. For the value part, we provide a value model gradient that admits a similar form as the well-known policy gradient [25, 26] (proof in Appendix D.1). The value model gradient is amenable to Monte Carlo approximation and can be computed exactly for finite MDPs.

Theorem 6.1 (Value model gradient)**.** *Suppose that the transition function P^M and reward function R^M of model M is parameterized by θ, the gradient of the value $V_1^{\pi,M}(s_1)$ w.r.t. the model is*

$$\nabla_\theta V_1^{\pi,M}(s_1) = \mathbb{E}_{\tau\sim\pi,M}\left[\sum_{h=1}^{H}\nabla_\theta \bar{R}^M(s_h,a_h) + \sum_{h=1}^{H-1} V_{h+1}^{\pi,M}(s_{h+1})\nabla_\theta \log P^M(s_{h+1}|s_h,a_h)\right], \tag{8}$$

where $\tau = (s_1, a_1, \ldots, s_H, a_H)$ is a trajectory, and $P^M(s_{h+1}|s_h,a_h)$ is the probability of s_{h+1} under distribution $P^M(s_h,a_h)$.

Nevertheless, the value model gradient suffers the same inefficiency just as the policy gradient [27] since the model gradient for a particular state-action pair is 0 whenever it is not visited under the current model and policy. The natural policy gradient proposed by [28] improves the efficiency of the policy gradient [29] by preconditioning the gradient with the inverse Fisher information matrix. We formalize the natural model gradient in a similar fashion.

Let D_k^R be normalized empirical state-action distribution for rewards $D_k^R(s,a) \propto \sum_{i=1}^{k-1}\sum_{h=1}^{H}\mathbb{I}(s_{ih}=s, a_{ih}=a)$. The normalized empirical state-action distribution for transitions is defined slightly differently as $D_k^P(s,a) \propto \sum_{i=1}^{k-1}\sum_{h=1}^{H-1}\mathbb{I}(s_{ih}=s, a_{ih}=a)$ since the transition pair in the H-th period is not observed by definition. Denoting the BOO objective in the k-th episode as $J_k = \lambda_k \log \Pr(M|\mathcal{H}_{k1}) + V_1^{\pi,M}(s_1)$, the natural model gradient is $\bar{\nabla}_\theta J_k = F_k^{-1}\nabla_\theta J_k$, where $F_k = \mathbb{E}_{\tau\sim\pi,M}[\frac{1}{H}\sum_{h=1}^{H} G^{\bar{R}}_{s_h,a_h} + \frac{1}{H-1}\sum_{h=1}^{H-1} F^P_{s_h,a_h}] + \mathbb{E}_{(s,a)\sim D_k^R}[F^R_{s,a}] + \mathbb{E}_{(s,a)\sim D_k^P}[F^P_{s,a}]$. Here, $F^P_{s,a} = \mathbb{E}_{s'\sim P^M(s_h,a_h)}[(\nabla_\theta \log P^M(s'|s_h,a_h))^2]$ and $F^R_{s,a} = \mathbb{E}_{r\sim R^M(s_h,a_h)}[(\nabla_\theta \log R^M(r|s_h,a_h))^2]$ are the Fisher information matrix of transition and reward distributions. The matrix $G^{\bar{R}}_{s,a} = \nabla_\theta z_R^\top \nabla^2_{z_R}\bar{R}^M(s,a)\nabla_\theta z_R$ is the generalized Gauss-Newton matrix [30], where the reward distribution is parameterized by $z_R \subseteq z$, and $z = f_\theta(s,a)$ is the output of the model function f_θ. Note that we have $G^{\bar{R}}_{s,a} = \mathbf{0}$ when the mean reward $\bar{R}^M(s,a)$ is linear in z. This is satisfied for most model classes, for example, those specified by Assumption 1. Therefore, commonly, F_k contains only the Fisher information matrix, and the natural gradient can be interpreted geometrically as the steepest descent direction in the information geometry induced by F_k instead of in the Euclidean geometry. Loosely speaking, the distance in the information geometry is measured by the square root of KL divergence between distributions rather than the ℓ_2 distance of parameters [31]. The natural gradient characterizes some global characteristics of the loss surface and therefore facilitates global convergence. We refer interested readers to [31] for a thorough discussion on the interpretation and implementation of the natural gradient.

Another intricacy of this problem is that the optimal solution of the BOO objective could change dramatically across the parameter space from episode to episode, which renders the optimization

extremely hard. A revealing fact is that a small change in the model could change the optimal policy dramatically, making the loss landscape unsmooth. Hence, we introduce an entropy-regularized optimization procedure, which starts with a high initial entropy regularization and gradually annealing. The entropy plays a role in smoothing the policy's loss landscape such that the optimal policy will not change drastically when the model changes. The smoothing effect is also observed previously in [32].

The entropy-regularized learning objective mimics the maximum entropy RL [33]:

$$\tilde{J}_k = J_k - \lambda_k \zeta \, \mathbb{E}_{\tau \sim \pi, M} \left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h(s_h) \| \hat{\pi}_h(s_h)) \right] = \lambda_k \log \Pr(M | \mathcal{H}_{k1}) + \tilde{V}_1^{\pi, M}(s_1), \tag{9}$$

where $\mathrm{KL}(p\|q)$ stands for the relative entropy, $\hat{\pi}_h$ is a prior policy ensuring π_h is absolutely continuous w.r.t. $\hat{\pi}_h$, the hyperparameter ζ controls the amount of entropy, and $\tilde{V}$ is the entropy-regularized value. Here, the entropy term is downscaled such that the influence of regularization diminishes with time. Consider the case that there exists a ϵ-optimal policy π^ϵ for any fixed model M such that, for some $b > 0$, $\mathbb{E}_{\tau \sim \pi^\epsilon, M} \left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h^\epsilon(s_h) \| \hat{\pi}_h(s_h)) \right] \leq b$ and $\sup_{\pi^*} V_1^{\pi^*, M}(s_1) - V_1^{\pi_\epsilon, M}(s_1) \leq \epsilon$. Then, the value of the maximum entropy policy $\tilde{\pi}^* = \arg\max_{\tilde{\pi}} V_1^{\tilde{\pi}, M}(s_1)$ satisfies that

$$\begin{aligned} V_1^{\tilde{\pi}^*, M}(s_1) &\geq V_1^{\tilde{\pi}^*, M}(s_1) - \lambda_k \zeta \, \mathbb{E}_{\tau \sim \tilde{\pi}^*, M} \left[\sum_{h=1}^{H} \mathrm{KL}(\tilde{\pi}_h^*(s_h) \| \hat{\pi}_h(s_h)) \right] \\ &\geq V_1^{\pi^\epsilon, M}(s_1) - \lambda_k \zeta \, \mathbb{E}_{\tau \sim \pi^\epsilon, M} \left[\sum_{h=1}^{H} \mathrm{KL}(\pi_h^\epsilon(s_h) \| \hat{\pi}_h(s_h)) \right] \\ &\geq \sup_{\pi^*} V_1^{\pi^*, M}(s_1) - \epsilon - \lambda_k \zeta b. \end{aligned} \tag{10}$$

If ϵ is sufficiently small, then the sub-optimality gap caused by entropy regularization will decrease at the same rate as the decay of value uncertainty, which ensures that the resulting policy covers multiple uncertain actions without detriment to the performance.

The entropy-regularized value is equivalently defined by the following Bellman backup,

$$\begin{aligned} &\tilde{V}_h^{\pi, M}(s_h) = \mathbb{E}_{a_h \sim \pi_h(s_h)}[\tilde{Q}_h^{\pi, M}(s_h, a_h)] - \lambda_k \zeta \mathrm{KL}(\pi_h(s_h) \| \hat{\pi}_h(s_h)), \, \forall h \in [H], \\ &\tilde{Q}_h^{\pi, M}(s_h, a_h) = \bar{R}^M(s_h, a_h) + \mathbb{E}_{s_{h+1} \sim P^M(s_h, a_h)} \left[\tilde{V}_{h+1}^{\pi, M}(s_{h+1}) \right], \, \forall h \in [H-1], \end{aligned} \tag{11}$$

where $\tilde{Q}_H^{\pi, M}(s_H, a_H) = \bar{R}^M(s_H, a_H)$. The optimization of the policy can be carried out by, for example, the maximum entropy actor-critic algorithm [33] or the soft actor-critic algorithm [34]. In finite MDPs, the optimal value is given by soft value iteration,

$$\tilde{V}_h^{\tilde{\pi}^*, M}(s_h) = \lambda_k \zeta \log \sum\nolimits_{a_h \in \mathcal{A}} \hat{\pi}(a_h | s_h) \exp(\tilde{Q}_h^{\tilde{\pi}^*, M}(s_h, a_h) / (\lambda_k \zeta)). \tag{12}$$

The maximum entropy policy $\tilde{\pi}^*$ is then derived from the optimal value as

$$\tilde{\pi}_h^*(s_h, a_h) = \hat{\pi}(a_h | s_h) \exp((\tilde{Q}_h^{\tilde{\pi}^*, M}(s_h, a_h) - \tilde{V}_h^{\tilde{\pi}^*, M}(s_h)) / (\lambda_k \zeta)). \tag{13}$$

The gradient of the entropy-regularized value w.r.t. the model is similar to that specified in Theorem 6.1, with the value replaced by the entropy-regularized value.

7 Conclusion

This paper proposes BOO as a generic model-based reinforcement learning algorithm. It enjoys sublinear regret for a variety of model classes. To optimize the BOO objective, we develop a natural model gradient that facilitates learning by optimizing the model in the space of distributions. We also propose to use an entropy-regularized objective that smooths the loss surface and makes BOO amenable to incremental implementation. However, to apply BOO in real-world reinforcement learning problems, it is essential to adapt the scaling parameter of BOO on a per-instance basis, which is an appealing direction for future work.

References

[1] Mohammed Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar, *Bayesian Reinforcement Learning: A Survey*. now, 2015.

[2] C. H. Papadimitriou and J. N. Tsitsiklis, "The Complexity of Markov Decision Processes," *Mathematics of Operations Research*, vol. 12, pp. 441–450, Aug. 1987.

[3] O. Madani, S. Hanks, and A. Condon, "On the undecidability of probabilistic planning and related stochastic optimization problems," *Artificial Intelligence*, vol. 147, pp. 5–34, July 2003.

[4] M. G. Azar, I. Osband, and R. Munos, "Minimax Regret Bounds for Reinforcement Learning," in *International Conference on Machine Learning*, pp. 263–272, PMLR, July 2017.

[5] S. R. Sinclair, S. Banerjee, and C. L. Yu, "Adaptive Discretization for Episodic Reinforcement Learning in Metric Spaces," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, pp. 1–44, Dec. 2019.

[6] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proceedings of Thirty Third Conference on Learning Theory*, pp. 2137–2143, PMLR, July 2020.

[7] V. Dani, T. Hayes, and S. Kakade, "Stochastic Linear Optimization Under Bandit Feedback," *21st Annual Conference on Learning Theory*, pp. 355–366, Jan. 2008.

[8] I. Osband and B. V. Roy, "Why is Posterior Sampling Better than Optimism for Reinforcement Learning?," in *International Conference on Machine Learning*, pp. 2701–2710, PMLR, July 2017.

[9] E. Kaufmann, O. Cappe, and A. Garivier, "On Bayesian Upper Confidence Bounds for Bandit Problems," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 592–600, PMLR, Mar. 2012.

[10] A. Tossou, C. Dimitrakakis, and D. Basu, "Near-optimal Bayesian Solution For Unknown Discrete Markov Decision Process," *arXiv preprint arXiv:1906.09114*, 2019.

[11] P. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains," *IEEE Transactions on Automatic Control*, vol. 27, pp. 137–146, Feb. 1982. Conference Name: IEEE Transactions on Automatic Control.

[12] X. Liu, P.-C. Hsieh, Y. H. Hung, A. Bhattacharya, and P. Kumar, "Exploration Through Reward Biasing: Reward-Biased Maximum Likelihood Estimation for Stochastic Multi-Armed Bandits," in *Proceedings of the 37th International Conference on Machine Learning*, pp. 6248–6258, PMLR, Nov. 2020. ISSN: 2640-3498.

[13] Y.-H. Hung, P.-C. Hsieh, X. Liu, and P. R. Kumar, "Reward-Biased Maximum Likelihood Estimation for Linear Stochastic Bandits," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 7874–7882, May 2021. Number: 9.

[14] A. Mete, R. Singh, X. Liu, and P. R. Kumar, "Reward Biased Maximum Likelihood Estimation for Reinforcement Learning," in *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, pp. 815–827, PMLR, May 2021. ISSN: 2640-3498.

[15] A. Mete, R. Singh, and P. R. Kumar, "Augmented RBMLE-UCB Approach for Adaptive Control of Linear Quadratic Systems," *arXiv:2201.10542 [cs, eess, math]*, Jan. 2022. arXiv: 2201.10542.

[16] R. J. Hyndman, "Computing and Graphing Highest Density Regions," *The American Statistician*, vol. 50, pp. 120–126, May 1996.

[17] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. CRC Press, third edition ed., Nov. 2013.

[18] R. J. Hyndman and Y. Fan, "Sample Quantiles in Statistical Packages," *The American Statistician*, vol. 50, pp. 361–365, Nov. 1996.

[19] M. J. A. Strens, "A Bayesian Framework for Reinforcement Learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, (San Francisco, CA, USA), pp. 943–950, Morgan Kaufmann Publishers Inc., June 2000.

[20] O. A. Adekola, *Asymptomatic posterior normality for stochastic processes.* PhD thesis, University of Surrey, Guildford, 1987.

[21] D. Russo and B. Van Roy, "Eluder Dimension and the Sample Complexity of Optimistic Exploration," in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013.

[22] D. Russo and B. Van Roy, "Learning to Optimize via Posterior Sampling," *Mathematics of Operations Research*, vol. 39, pp. 1221–1243, Nov. 2014.

[23] C. Jin, Q. Liu, and S. Miryoosefi, "Bellman Eluder Dimension: New Rich Classes of RL Problems, and Sample-Efficient Algorithms," in *Advances in Neural Information Processing Systems*, May 2021.

[24] Z. Yang, C. Jin, Z. Wang, M. Wang, and M. I. Jordan, "On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces," *arXiv:2011.04622 [cs, math, stat]*, Dec. 2020. arXiv: 2011.04622.

[25] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, pp. 229–256, May 1992.

[26] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy Gradient Methods for Reinforcement Learning with Function Approximation," in *Advances in Neural Information Processing Systems* (S. Solla, T. Leen, and K. Müller, eds.), vol. 12, MIT Press, 2000.

[27] S. Kakade and J. Langford, "Approximately Optimal Approximate Reinforcement Learning," in *In Proc. 19th International Conference on Machine Learning*, pp. 267–274, 2002.

[28] S. M. Kakade, "A Natural Policy Gradient," in *Advances in Neural Information Processing Systems*, vol. 14, MIT Press, 2002.

[29] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes," in *Proceedings of Thirty Third Conference on Learning Theory*, pp. 64–66, PMLR, July 2020.

[30] N. N. Schraudolph, "Fast Curvature Matrix-Vector Products for Second-Order Gradient Descent," *Neural Computation*, vol. 14, pp. 1723–1738, July 2002. Conference Name: Neural Computation.

[31] J. Martens, "New Insights and Perspectives on the Natural Gradient Method," *Journal of Machine Learning Research*, vol. 21, no. 146, pp. 1–76, 2020.

[32] Z. Ahmed, N. L. Roux, M. Norouzi, and D. Schuurmans, "Understanding the Impact of Entropy on Policy Optimization," in *Proceedings of the 36th International Conference on Machine Learning*, pp. 151–160, PMLR, May 2019. ISSN: 2640-3498.

[33] S. Levine, "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review," *arXiv:1805.00909 [cs, stat]*, May 2018.

[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," in *International Conference on Machine Learning*, pp. 1861–1870, PMLR, July 2018.

[35] K. Huang, S. M. Kakade, J. D. Lee, and Q. Lei, "A Short Note on the Relationship of Information Gain and Eluder Dimension," *arXiv:2107.02377 [cs, math, stat]*, July 2021.

[36] M. Brookes, "The Matrix Reference Manual," 2020.

[37] I. Osband and B. Van Roy, "Model-based Reinforcement Learning and the Eluder Dimension," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[38] A. Maurer, "A bound on the deviation probability for sums of non-negative random variables.," *JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only]*, vol. 4, no. 1, pp. Paper No. 15, 6 p., electronic only–Paper No. 15, 6 p., electronic only, 2003. Publisher: Victoria University, School of Communications and Informatics.

[39] I. Osband, D. Russo, and B. Van Roy, "(More) Efficient Reinforcement Learning via Posterior Sampling," in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] No relevant potential negative impacts is identified.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]