

Variational Bayesian Reinforcement Learning with Regret Bounds

Brendan O'Donoghue

DeepMind

bodonoghue@google.com

July 25, 2018

Abstract

We consider the exploration-exploitation trade-off in reinforcement learning and we show that an agent imbued with a risk-seeking utility function is able to explore efficiently, as measured by regret. The parameter that controls how risk-seeking the agent is can be optimized exactly, or annealed according to a schedule. We call the resulting algorithm K-learning and show that the corresponding K-values are optimistic for the expected Q-values at each state-action pair. The K-values induce a natural Boltzmann exploration policy for which the ‘temperature’ parameter is equal to the risk-seeking parameter. This policy achieves an expected regret bound of $\tilde{O}(L^{3/2}\sqrt{SAT})$, where L is the time horizon, S is the number of states, A is the number of actions, and T is the total number of elapsed time-steps. This bound is only a factor of L larger than the established lower bound. K-learning can be interpreted as mirror descent in the policy space, and it is similar to other well-known methods in the literature, including Q-learning, soft-Q-learning, and maximum entropy policy gradient, and is closely related to optimism and count based exploration methods. K-learning is simple to implement, as it only requires adding a bonus to the reward at each state-action and then solving a Bellman equation. We conclude with a numerical example demonstrating that K-learning is competitive with other state-of-the-art algorithms in practice.

Contents

1	Introduction and related work	3
2	Preliminaries	5
2.1	Epistemic uncertainty vs Aleatoric uncertainty	5
2.2	Utility functions	5
2.3	Cumulant generating function	6
2.4	Bounding the value of a max	7
2.5	Variational representation	8
2.6	Saddle point problem	8
3	Stochastic multi-armed bandits	9
3.1	Gap-dependent bound	13
3.2	Numerical experiments	13
4	Markov decision processes	15
4.1	The case with known dynamics	17
5	K-learning	18
5.1	The case with known dynamics	19
5.2	The case with unknown dynamics	19
5.2.1	Greedy formulation	22
5.3	Formulation as a convex optimization problem	23
5.4	Numerical example	25
5.5	Connections with other methods	26
6	Conclusions	28
7	Acknowledgments	28
8	Appendix	36
8.1	Greedy K-learning for Bandits	36
8.2	Greedy K-learning for MDPs	36
8.3	The perspective preserves strict convexity	36
8.4	Conic representation of perspective of log-sum-exp	37

1 Introduction and related work

We consider the reinforcement learning problem, whereby an agent interacts with an environment in an episodic manner and attempts to maximize its return given the information it obtains from its interactions [102, 16], and where the environment is a Markov decision process (MDP) [84]. In this paper we consider the Bayesian case, where the agent has some prior information, and as it gathers data it updates its posterior beliefs about the environment. In this setting the agent is faced with the choice between taking well understood actions or exploring the environment to determine the value of other actions which might lead to a higher return. This dilemma is called the *exploration-exploitation* trade-off, and exploring efficiently is an important area of research. One way to measure how well an agent explores is a quantity called *regret*, which measures how sub-optimal the rewards the agent has received are so far, relative to the (unknown) optimal policy [20]. In the Bayesian case the natural quantity to consider is the Bayesian expected regret, which is the expected regret under the agents prior information [29]. Known lower bounds to the regret exist, which is to say that no algorithm can do better than the bound on every problem. In this paper we present a new policy that achieves a Bayesian expected regret close to the lower bound, and which matches the regret of other methods in the literature.

There are several other algorithms that get provable regret bounds. In the frequentist case regret bounds hold with high probability or in expectation over realizations of problems in a given class, and do not assume any prior information [39]. Most of these techniques rely on *optimism in the face of uncertainty*, whereby the agent searches for a value function that is a high-probability upper bound on the true value function, then follows the greedy policy for that value function [57, 103, 100]. This process may involve, for example, constructing an upper-bound on the reward function using a concentration inequality such as Hoeffding’s lemma, and then searching over the space of transition functions within some confidence set for the one that gives the largest value function. The algorithms UCRL and UCRL2 [8, 36] fall into this category for the MDP case, and UCB1 for the bandit case [6]. Various extensions that obtain better regret bounds exist and are an active area of research [10, 45].

In the Bayesian case the optimal policy can be formulated using *belief states*, but this is intractable for all but small problems [29]. In the multi-armed bandit case it can be formulated using Gittins indices [30], which converts a bandit problem into an MDP which can be solved for the optimal policy. Approximations to the optimal Bayesian policy exist, one of the most successful being Thompson sampling, also known as probability matching [101, 105]. In Thompson sampling the agent samples from the posterior over value functions and acts greedily with respect to that sample [74, 78, 47, 77], and it can be shown that this strategy yields both Bayesian and frequentist regret bounds (under certain assumptions) [3]. In practice, maintaining a posterior over value functions is intractable, and so instead the agent maintains the posterior over MDPs, and at each episode an MDP is sampled from this posterior, the value function for that MDP solved for, and the policy is greedy with respect to that value function. The benefit of using a Bayesian approach include the ability to incorporate prior information, as well as not requiring a computationally difficult search over transition functions. However, due to the nature of resampling they are practical only for very small problems, though attempts have been made to extend them [73, 64].

Though we do not consider it here it is worth mentioning the PAC-MDP (probably approximately correct for MDPs) framework, which is an alternative to regret to quantify the performance of reinforcement learning algorithms. An algorithm is said to be PAC-MDP if it is guaranteed to have close to optimal performance with high probability after a number of time-steps that is polynomial in the quantities that describe the problem [18, 39, 23, 41, 98, 111, 99]. However, these strategies generally suffer from *linear* regret, since many PAC-MDP algorithms explicitly halt learning once an almost-optimal policy has been found. In fact,

it can be shown that a PAC-MDP algorithm cannot have sub-linear expected regret [24, Thm. 1].

In this paper we propose a practical Bayesian algorithm that attains an expected regret upper bounded by $\tilde{O}(L^{3/2}\sqrt{SAT})$ where L is the time horizon, S is the number of states, A is the number of actions per state, and T is the total number of elapsed time-steps (where \tilde{O} ignores logarithmic factors). This matches the bounds for several other Bayesian methods in the literature, see *e.g.*, [78]. Our regret bound is within a factor of L of the known minimax lower bound of $\Omega(\sqrt{LSAT})$ (although strictly speaking these bounds are not comparable due to the assumptions we make in order to derive our bound). Our technique can be interpreted as optimism in the face of *Bayesian* uncertainty, since we propagate values corresponding to a risk-seeking utility function, which are optimistic for the true Q-values. Our technique is similar in spirit to BEB (Bayesian Exploration Bonus) [44], BOLT (Bayesian Optimistic Local Transitions) [4], and Bayes-UCB (Bayesian Upper Confidence Bounds) [40], which use Bayesian information to incorporate optimism into the values. Both BEB and BOLT are PAC-BAMDP, which means they produce policies that are close to the optimal *Bayesian* policy. This is a weaker notion than PAC-MDP which is with respect to the optimal *instance* policy. Bayes-UCB achieves a frequentist regret bound for binary stochastic bandits.

The Von Neumann-Morgenstern utility theorem states that any *rational* agent must behave as though it is maximizing the expected value of a *utility function*, where rationality is defined with respect to a set of axioms [107]. In the reinforcement literature there is a lot of work on exploration strategies such as optimism, curiosity, intrinsic motivation, surprise, novelty seeking, *etc.*, [80, 33, 43, 49, 56, 21]. Assuming these agents are rational, then in the language of utility theory most of these prior works are equivalent to giving the agent a non-linear utility function. In this work we imbue an agent with a particular *risk-seeking* utility function, where risk is taken to be the uncertainty that the agent has about its own estimates of the values of each state-action. Any increasing convex function could be used as a risk-seeking utility, however only the exponential utility function has the *shift-invariance* property which we will use in order to derive a Bellman-like equation for the values under the risk-seeking utility.

The update rule we derive is similar to that used in ‘soft’ Q-learning (so-called since the ‘hard’ max is replaced with a soft-max) [9, 28, 31, 58, 86]. The soft-max operator can be justified by viewing reinforcement learning as probabilistic inference. In this view the likelihood that a sequence of actions is optimal is proportional to the exponential of the total reward accumulated along the trajectory, and the goal is to infer the optimal policy; see [46] for a recent survey. These approaches are very closely related to maximum entropy reinforcement learning, which adds an entropy regularization ‘bonus’ to prevent early convergence to deterministic policies and thereby encourages exploration, though there are no guarantees on regret [112, 113, 54, 63, 2]. In our work the soft-max operator and entropy regularization arise naturally from the view of the agent as maximizing a risk-seeking utility. Furthermore, in contrast to these other approaches, the entropy regularization is not a fixed hyper-parameter but something we explicitly control (or optimize over) in order to derive a regret bound.

Paper organization. In the next section we lay out some preliminary identities and notation that we will use for the remainder of the paper. In the following section we consider the K-learning policy applied to the stochastic multi-armed bandit case and we derive a Bayesian expected regret bound that matches the lower bound (up to log factors). In the following section we extend the analysis to Markov Decision Processes where we have the cumulant generating function for the full posterior over the Q-values (though we do not discuss how to calculate those posteriors). This finally motivates the introduction of *K*-learning, where the solutions to a Bellman equation approximate the cumulant generating function of the true posterior, which may be intractable to compute exactly. We introduce *K*-learning in two parts, first we consider the case where only the reward function is unknown, and finally extend to the full case where both the reward and

dynamics are unknown. We derive a Bayesian expected regret bound for the full problem that is within a factor of the episode length from the known lower bound.

2 Preliminaries

In this section we cover some basic concepts that we will use throughout this paper. We consider all random variables to be defined with respect to a probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

2.1 Epistemic uncertainty vs Aleatoric uncertainty

In this paper we are mostly concerned with the uncertainty associated with estimating a parameter using finite data, since it is this type of uncertainty that is useful for exploration [64]. We assume we have access to a prior over possible values of the parameter, which allows us to do Bayesian inference. This type of uncertainty is sometimes referred to as *epistemic* uncertainty, or model uncertainty. Epistemic uncertainty is distinct from the so-called *aleatoric* uncertainty or risk, which is the natural randomness of the process we are observing [96, 104, 14]. This difference can be made clear by an example. Consider flipping a biased coin and attempting to learn from the flips the probability of getting a head, denoted p . After any number of flips we have a posterior over possible values of p , which in the limit will concentrate around the true value. However, the randomness of the process of flipping coins will not change as we accumulate data. So in this case the epistemic uncertainty, the uncertainty about our estimate of p , decreases with more data, whereas the aleatoric uncertainty, the inherent randomness of the coin flipping process, is fixed.

2.2 Utility functions

A utility function $u : \mathbb{R} \rightarrow \mathbb{R}$ measures the satisfaction or benefit that an agent derives from a particular outcome. A rational agent will seek to maximize its expected utility, $\mathbf{E}u(X)$ for some random payoff X [107]. If u is concave then it is referred to as *risk-averse*, in that an agent that maximizes a risk-averse utility function will prefer more predictable payoffs. Risk-averse utilities are commonly used in finance and insurance, where in those contexts risk refers to the randomness of the returns, *i.e.*, the aleatoric uncertainty. If u is convex then it is referred to as *risk-seeking*, in that an agent with a risk-seeking utility will prefer less predictable payoffs. If u is the identity (or, more generally, affine) then the agent is *risk-neutral*, *i.e.*, it only cares about the expected value of the payoff and has no preference one way or the other for the risk. In reinforcement learning it is generally assumed that the agent is risk-neutral, *i.e.*, that it is attempting to maximize the (possibly discounted) expected sum of future returns. However, in this manuscript we show that when the agent has epistemic uncertainty about the environment, the usual case, then a risk-seeking utility can be used to explore efficiently. Put another way, an agent imbued with a risk-seeking utility function will prefer outcomes with greater epistemic uncertainty, which will cause the agent to explore parts of the state space it hasn't been to before.

An important concept in the utility literature is that of the *certainty equivalent value*, which is the amount of guaranteed payoff that an agent considers as equally desirable to an uncertain payout. For an invertible utility function u and random payoff X it has the following form

$$C_X = u^{-1}(\mathbf{E}u(X)).$$

If u is convex (*i.e.*, the agent is risk-seeking) then Jensen's inequality implies that $C_X \geq \mathbf{E}(X)$; the opposite holds for concave u . In other words, for convex u the certainty equivalent value is *optimistic* for the expected value of a random payoff.

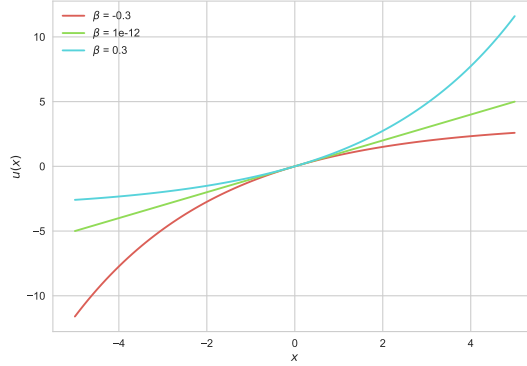


Figure 1: The exponential utility function for different choices of β .

The most important relationship in control and reinforcement learning is the *Bellman equation*, which equates the value of the current state to the immediate reward plus the value of the next state. The Bellman equation assumes there is no uncertainty in the quantities (or alternatively, that the agent is risk-neutral with respect to the uncertainty). Under a risk-sensitive utility we would like to maintain the ability to decompose the value at a state into the immediate value plus the value at the next state. Formally, we would like a utility function that for independent random variables X and Y satisfies the following

$$C_{X+Y} = C_X + C_Y.$$

It turns out that the *exponential utility function* is the *unique* utility function that satisfies this property (the identity utility is a special case) [1, 81, 34, 85], which, up to affine transformation, is given by

$$u(X) = (1/\beta)(\exp(\beta X) - 1).$$

where β is a parameter that controls how risk-sensitive the agent is; see fig. 1. In this case the certainty equivalent value takes the form given by

$$C_X(\beta) = (1/\beta) \log \mathbf{E} \exp(\beta X). \quad (1)$$

For $\beta < 0$ this utility function is risk-averse and has been studied extensively in the literature on safety and risk-averse reinforcement learning [35, 83, 52, 53, 48, 50]. It becomes risk-neutral in the limit of $\beta \rightarrow 0$. In this paper we shall consider the case where $\beta > 0$, in which case the utility is risk-seeking, a property we shall exploit for efficient exploration.

2.3 Cumulant generating function

A fundamental object of statistics that we make use of in this work is the *cumulant generating function* of a random variable [42], which for random variable X is defined as

$$K_X(\beta) = \log \mathbf{E} \exp(X\beta), \quad (2)$$

i.e., the log of the moment generating function. The cumulant generating function (if it exists) is convex in β [17]. An alternative way to write the cumulant generating function is a series of the cumulants

$$K_X(\beta) = \sum_{n=1}^{\infty} \kappa_n \beta^n / n!$$

where κ_n is the n -th cumulant, the first and second cumulant are the mean and variance of X respectively. With this we can rewrite the value of the exponential utility function in eq. (1) as

$$C_X(\beta) = (1/\beta)K_X(\beta).$$

2.4 Bounding the value of a max

When we come to the reinforcement learning problem we will consider an agent in a Markov Decision Process (MDP) interested in maximizing its return. However, the agent will not know which actions lead to the maximum return, instead it may have a Bayesian posterior, or an approximate posterior, over the possible returns of the actions it can take at any time-step. In order to propagate the value of the current state backwards to earlier states (*i.e.*, dynamic programming) the agent will need to calculate the certainty equivalent value of the random variable corresponding to the max over its possible actions, under the *epistemic* uncertainty it is faced with. Though we cannot calculate this value exactly in general, we can provide a bound.

Lemma 1. *Given a collection of random variables X_i , $i = 1, \dots, k$ let $Y = \max_i X_i$, the cumulant generating function for Y satisfies*

$$\log \sum_i \exp K_{X_i}(\beta) - \log k \leq K_Y(\beta) \leq \log \sum_i \exp K_{X_i}(\beta). \quad (3)$$

Proof. From the definition of the cumulant generating function

$$K_Y(\beta) = \log \mathbf{E} \exp \beta Y = \log \mathbf{E} \max_i \exp \beta X_i$$

and the lemma follows since the max over a collection of nonnegative numbers is bounded above by the sum and bounded below by the mean. \square

Corollary 1. *Under the same conditions as the previous lemma*

$$\max_i K_{X_i}(\beta) \leq K_Y(\beta) \leq \max_i K_{X_i}(\beta) + \log k. \quad (4)$$

Proof. Which holds since for any collection of numbers x_i , $i = 1, \dots, k$, we have $\max_i x_i \leq \log \sum_{i=1}^k \exp x_i \leq \max x_i + \log k$. \square

Based on these lemmas we can bound the expected value of the max of a collection of random variables. Let $Y = \max_i X_i$, then since the certainty equivalent value is optimistic for the mean and using the bound from (3)

$$\mathbf{E} \max_i X_i \leq C_Y(\beta) = (1/\beta)K_Y(\beta) \leq (1/\beta) \log \sum_{i=1}^k \exp K_{X_i}(\beta), \quad (5)$$

and furthermore the problem of finding the tightest bound by minimizing over $\beta \geq 0$ is a convex optimization problem in the variable $\tau = 1/\beta$.

Lemma 2. *Let K_{X_i} denote the cumulant generating function for random variable X_i , then the function $g(v, \tau) = \tau \log \sum_{i=1}^k \exp K_{X_i}(v/\tau)$ is convex in v and τ jointly.*

Proof. Recall that K_{X_i} is convex for each i , and log-sum-exp is a convex non-decreasing function, so the composition h given by $h(v) = \log \sum_{i=1}^k \exp K_{X_i}(v)$ is convex in v . Finally, note that g is the perspective of h and therefore convex in v and τ jointly [17]. \square

Corollary 2. *The function $\tau \log \sum_{i=1}^k \exp K_{X_i}(1/\tau)$ is convex, similarly the function $\tau K_X(1/\tau)$ is convex.*

2.5 Variational representation

Convex duality is a rich source of variational (*i.e.*, optimization based) representations of problems [108]. Here we use this principle to derive a policy based on our bound on the value of the max under the exponential utility.

Lemma 3. *Given a collection of random variables X_i , $i = 1, \dots, k$, we have for any $\beta > 0$*

$$\log \sum_{i=1}^k \exp K_{X_i}(\beta) = \max_{\pi \in \Pi} \left(\sum_{i=1}^k \pi_i K_{X_i}(\beta) + H(\pi) \right) \quad (6)$$

where Π is the probability simplex and H denotes the entropy, *i.e.*, $H(\pi) = -\sum_{i=1}^k \pi_i \log \pi_i$ [22]. The maximum is achieved by the policy

$$\pi_i^* = \exp K_{X_i}(\beta) / \sum_{j=1}^k \exp K_{X_j}(\beta). \quad (7)$$

Proof. This comes from taking the Legendre transform of negative entropy term (equivalently, log-sum-exp and negative entropy are convex conjugates [17, Example 3.25]). The fact that π^* achieves the maximum is readily verified by substitution. \square

The right hand side of equation (6) has the form of negative Helmholtz free energy, and the optimal policy is therefore minimizing this energy. In the sequel we shall show that following policy π^* yields a regret bound close to the known lower bound. This variational policy has strong connections to the literature on decision making under model uncertainty or under information constraints, also called *bounded rationality*, see, *e.g.*, [70, 72, 68, 69, 71].

2.6 Saddle point problem

Consider a collection of *non-degenerate* random variables X_i for $i = 1 \dots, k$ and let

$$\mathcal{L}(\tau, \pi) = \sum_{i=1}^k \pi_i \tau K_{X_i}(1/\tau) + \tau H(\pi).$$

Note that \mathcal{L} is differentiable, strictly convex in τ , and strictly concave in π for $\tau > 0$. To see strict convexity note that since the X_i are non-degenerate it implies that the K_{X_i} are strictly convex [37, Thm. 2.3] and the perspective of a function preserves strict convexity (see appendix). We can rewrite the problem of finding the tightest bound on the expected value of the max of k random variables as the following saddle point problem

$$\mathbf{E} \max_i X_i \leq \min_{\tau \geq 0} \max_{\pi \in \Pi} \mathcal{L}(\tau, \pi),$$

and from the previous discussion we know this is achieved by $\tau^* = \operatorname{argmin}_{\tau \geq 0} \left(\tau \log \sum_{i=1}^k \exp K_{X_i}(1/\tau) \right)$ and $\pi^* \propto \exp K_{X_i}(1/\tau^*)$, which are unique.

Since \mathcal{L} is differentiable, convex-concave, and Π is compact, it satisfies a *strong max-min* or *saddle-point* property [17, Exercise 3.14][106, 95], *i.e.*,

$$\max_{\pi \in \Pi} \mathcal{L}(\tau^*, \pi) = \mathcal{L}(\tau^*, \pi^*) = \min_{\tau \geq 0} \mathcal{L}(\tau, \pi^*). \quad (8)$$

This observation brings us to the next lemma.

Lemma 4. *Given a collection of non-degenerate random variables X_i , $i = 1, \dots, k$, then there exists a unique τ^* that satisfies*

$$\tau^* = \operatorname{argmin}_{\tau \geq 0} \left(\tau \log \sum_{i=1}^k \exp K_{X_i}(1/\tau) \right) = \operatorname{argmin}_{\tau \geq 0} \left(\sum_{i=1}^k \pi_i^* \tau K_{X_i - \mathbf{E}(X_i)}(1/\tau) + \tau H(\pi^*) \right) \quad (9)$$

where $\pi_i^* \propto \exp K_{X_i}(1/\tau^*)$ and $K_{X_i - \mathbf{E}(X_i)}$ is the cumulant generating function of the centered posterior over X_i .

Proof. This follows since \mathcal{L} satisfies the strong min-max property (8)

$$\begin{aligned} \tau^* &= \operatorname{argmin}_{\tau \geq 0} \left(\tau \log \sum_{i=1}^k \exp K_{X_i}(1/\tau) \right) \\ &= \operatorname{argmin}_{\tau \geq 0} \left(\sum_{i=1}^k \pi_i^* \tau K_{X_i}(1/\tau) + \tau H(\pi^*) \right) \\ &= \operatorname{argmin}_{\tau \geq 0} \left(\sum_{i=1}^k \pi_i^* \mathbf{E}(X_i) + \sum_{i=1}^k \pi_i^* \tau K_{X_i - \mathbf{E}(X_i)}(1/\tau) + \tau H(\pi^*) \right) \\ &= \operatorname{argmin}_{\tau \geq 0} \left(\sum_{i=1}^k \pi_i^* \tau K_{X_i - \mathbf{E}(X_i)}(1/\tau) + \tau H(\pi^*) \right). \end{aligned}$$

□

The above lemma allows us to derive the following bound

$$\mathbf{E} \max_i X_i \leq \sum_{i=1}^k \pi_i^* \mathbf{E}(X_i) + \min_{\tau \geq 0} \left(\sum_{i=1}^k \pi_i^* \tau K_{X_i - \mathbf{E}(X_i)}(1/\tau) + \tau H(\pi^*) \right).$$

In the literature on online learning the term inside the minimization is sometimes referred to as the *mixability gap* [26] (generally used in a non-Bayesian framework). In our case we can explicitly find the τ^* that minimizes the mixability gap.

This insight will be useful later when we come to talk about regret and show that the right hand side of (9) is an upper bound on the Bayesian expected regret, this lemma will show us that the choice of τ that provides the tightest bound on the expected value of the max of a collection of random variables also provides the tightest upper bound on the expected regret.

3 Stochastic multi-armed bandits

In this section we prove a Bayesian expected regret bound for our policy applied to stochastic multi-armed bandits. Although bandits are not the main focus of this paper the analysis tools we use here will be useful in the sequel. In the stochastic multi-armed bandit setting an agent selects an ‘arm’ at each time-step and receives a corresponding reward; the goal is to identify the arm that provides the best payout in expectation. In this case our random variables are the mean payout of each arm, and we assume we have a posterior over each arm conditioned on the information we have received from previous actions and the prior distribution. Many algorithms exist that achieve low regret for bandits, see the survey [19], in particular the MOSS algorithm is known to achieve the optimal regret lower bound, up to constants, [5]. In the Bayesian case Thompson sampling also achieves close to the regret bound [87, 88], as does Bayes-UCB which aims to combine Bayesian methods and optimism [40].

A stochastic multi-armed bandit problem is defined by the pair $\{A, R\}$, where $A \in \mathbb{N}$ is the number of arms (actions) available to the agent to pull and $R(a)$ is a probability distribution over rewards for each

$a \in \{1, \dots, A\}$. At each time-step t the agent selects an arm $a_t \in \{1, \dots, A\}$ and receives a reward $r_t \sim R(a_t)$. Initially the agent does not know the distribution of the rewards, but must explore the arms to learn about them. We consider the Bayesian case, where the *mean* payoff of each arm is sampled from a known prior Φ_i , *i.e.*, $\mu_i \sim \Phi_i$ for each $i = 1, \dots, A$. The agent observes the reward it receives at each time-step and uses that information to perform a Bayesian update of its posterior beliefs over μ_i for the selected action i . The goal of the agent is to maximize its long-term reward, which, loosely speaking, means the agent must determine which arm has the highest expected reward and select that most frequently.

Let \mathcal{F}_t denote the sigma-algebra generated by all the history *before* time t (*i.e.*, it does not include the observations at time t), and we define $\mathcal{F}_1 = \emptyset$. Since we are interested in agents that accumulate information over time, we will require the *conditional* cumulant generating function of random variables conditioned on \mathcal{F}_t , which for random variable X is defined as

$$K_X^t(\beta) = \log \mathbf{E}(\exp(\beta X) | \mathcal{F}_t). \quad (10)$$

Consider the case where the agent is using the policy defined as

$$\pi_{it} = \exp K_{\mu_i}^t(\beta_t) / \sum_{j=1}^n \exp K_{\mu_j}^t(\beta_t). \quad (11)$$

We show that for a particular schedule of β_t this policy achieves an expected regret bound close to the known lower bound. The Bayesian expected regret up to time T is given by

$$\mathbf{E} R(T) = \mathbf{E} \sum_{t=1}^T (\max_i \mu_i - r_t),$$

where r_t is the reward obtained by the policy at time t , and the expectation operator is with respect to the mean reward μ_i of each arm being drawn from the prior and the randomness of the rewards and policy.

Applying the tower property of conditional expectation and using lemma 3 we have

$$\begin{aligned} \mathbf{E} R(T) &= \mathbf{E} \sum_{t=1}^T \mathbf{E}(\max_i \mu_i - r_t | \mathcal{F}_t) \\ &\leq \mathbf{E} \sum_{t=1}^T (1/\beta_t) \mathbf{E} \left(\sum_{i=1}^A \pi_{it} K_{\mu_i}^t(\beta_t) + H(\pi_t) - \beta_t r_t | \mathcal{F}_t \right), \end{aligned} \quad (12)$$

which holds for any sequence of β_t , $t = 1, \dots, T$. If we denote by $\sigma(\mu_i)$ the sigma-algebra generated by μ_i , then the tower property of condition expectation tells us that for any i

$$\mathbf{E}(r_i | \mathcal{F}_t) = \mathbf{E}(\mathbf{E}(r_i | \mathcal{F}_t \cup \sigma(\mu_i)) | \mathcal{F}_t) = \mathbf{E}(\mu_i | \mathcal{F}_t), \quad (13)$$

from which we obtain

$$\mathbf{E}(r_t | \mathcal{F}_t) = \mathbf{E} \left(\sum_{i=1}^A \pi_{it} r_i | \mathcal{F}_t \right) = \sum_{i=1}^A \pi_{it} \mathbf{E}(r_i | \mathcal{F}_t) = \sum_{i=1}^A \pi_{it} \mathbf{E}(\mu_i | \mathcal{F}_t),$$

since π_t is \mathcal{F}_t -measurable. Substituting in we obtain

$$\mathbf{E} R(T) \leq \mathbf{E} \sum_{t=1}^T (1/\beta_t) \left(\sum_{i=1}^A \pi_{it} K_{\mu_i - \mathbf{E}(\mu_i | \mathcal{F}_t)}^t(\beta_t) + H(\pi_t) \right), \quad (14)$$

where $K_{\mu_i - \mathbf{E}(\mu_i | \mathcal{F}_t)}^t$ is the cumulant generating function of the centered posterior over μ_i at time t . Since this holds for any sequence of β_t we can minimize this over β_t

$$\mathbf{E} R(T) \leq \mathbf{E} \sum_{t=1}^T \min_{\beta_t} (1/\beta_t) \left(\sum_{i=1}^A \pi_{it} K_{\mu_i - \mathbf{E}(\mu_i | \mathcal{F}_t)}^t(\beta_t) + H(\pi_t) \right), \quad (15)$$

and the optimal β_t is given by

$$\beta_t^* = \underset{\beta}{\operatorname{argmin}} (1/\beta) \log \sum_{i=1}^A \exp K_{\mu_i}^t(\beta) \quad (16)$$

by lemma 4. If we can derive a regret bound for any fixed sequence of β_t then it implies that the same regret bound holds for β_t^* . To do so we make the following standard assumption:

Assumption 1. *The posterior over the value of each arm is sub-Gaussian, and concentrates with data at least as fast as a Gaussian, i.e.,*

$$\mathbf{E}(\exp \beta(\mu_i - \mathbf{E}(\mu_i | \mathcal{F}_t)) | \mathcal{F}_t) \leq \exp(\beta^2 \sigma^2 / 2n_{it})$$

for some $\sigma \geq 0$ and where n_{it} is the number of times we have pulled arm i after t time-steps.

For example, if the arms rewards are Gaussian distributed, and the prior is Gaussian, then the assumption is satisfied. If the reward is categorical and the prior is Dirichlet then the posterior is also Dirichlet which is sub-Gaussian [51].

With this assumption we have

$$\mathbf{E} R(T) \leq \mathbf{E} \sum_{t=1}^T \sum_{i=1}^A (\pi_{it} \sigma^2 \beta_t / 2n_{it} + H(\pi_t) / \beta_t). \quad (17)$$

Now all that remains is to bound the two terms in the sum. To bound the first term we will require the following lemma.

Lemma 5. *Consider a process that at each time t selects a single index I_t from $\{1, \dots, A\}$ with probability $\pi_{I_t t}$. Let n_{it} denote the count of the number of times index i has been selected up to time t . Then*

$$\sum_{t=1}^T \sum_{i=1}^A \pi_{it} / n_{it} \leq A(1 + \log T).$$

Proof. This follows from a straightforward application of the pigeonhole principle,

$$\begin{aligned} \sum_{t=1}^T \sum_{i=1}^A \pi_{it} / n_{it} &= \sum_{t=1}^T \mathbf{E}_{I_t \sim \pi_t} 1/n_{I_t} \\ &= \mathbf{E}_{I_0 \sim \pi_0, \dots, I_T \sim \pi_T} \sum_{t=1}^T 1/n_{I_t} \\ &= \mathbf{E}_{I_0 \sim \pi_0, \dots, I_T \sim \pi_T} \sum_{i=1}^A \sum_{t=1}^{n_{iT}} 1/t \\ &\leq \sum_{i=1}^A \sum_{t=1}^T 1/t \\ &\leq A(1 + \log T), \end{aligned}$$

where the last inequality follows since $\sum_{t=1}^T 1/t \leq 1 + \int_{t=1}^T 1/t = 1 + \log(T)$. \square

With this we are ready to prove a regret bound for this strategy.

Theorem 1. *The K-learning algorithm 1 achieves Bayesian expected regret bound*

$$\mathbf{E} R(T) \leq 2\sigma \sqrt{TA \log A(1 + \log T)}.$$

Proof. The bound in eq. (17) holds for any sequence of β_t so we can set

$$\beta_t = \sqrt{\frac{4t \log A}{\sigma^2 A(1 + \log t)}},$$

and with this choice we can bound the two terms in the regret bound separately. Firstly, we have

$$\begin{aligned} \sum_{t=1}^T H(\pi_t)/\beta_t &\leq (\sigma/2) \sqrt{A \log A(1 + \log T)} \sum_{t=1}^T 1/\sqrt{t} \\ &\leq \sigma \sqrt{TA \log A(1 + \log T)}, \end{aligned} \tag{18}$$

since $\sum_{t=1}^T 1/\sqrt{t} \leq \int_{t=0}^T 1/\sqrt{t} = 2\sqrt{T}$. To bound the remaining term we use lemma 5

$$\begin{aligned} \sum_{t=1}^T \sum_{i=1}^A \sigma^2 \beta_t \pi_{it} / 2n_{it} &\leq \sigma^2 \beta_T A(1 + \log T)/2 \\ &= \sigma \sqrt{TA \log A(1 + \log T)}. \end{aligned}$$

The result is obtained by summing these two bounds. \square

The above theorem matches (up to logarithmic factors) the lower bound on Bayesian expected regret, since it can be shown (e.g., [19, Thm. 3.5]) that there exist prior distributions such that for any algorithm one has

$$\mathbf{E} R(T) \geq c\sqrt{TA},$$

for some $c > 0$.

Corollary 3. *The K-learning algorithm where at each time $\beta_t = \operatorname{argmin}_{\beta} (1/\beta) \log \sum_{i=1}^k \exp K_{\mu_i}^t(\beta)$ attains a Bayesian expected regret bound at least as good as the previous theorem.*

If we consider the case where $\beta_t = \beta$ for some fixed β then using the bounds derived above we can rewrite eq. (17) as

$$\mathbf{E} R(T) \leq \frac{\beta \sigma^2 A(1 + \log T)}{2} + \frac{T \log A}{\beta}, \tag{19}$$

and minimizing over β yields

$$\mathbf{E} R(T) \leq \sigma \sqrt{2TA \log A(1 + \log T)},$$

which is a slightly better bound than the above theorem, however this choice of β requires knowledge of T in advance and so we lose the anytime nature. Writing the regret as in eq. (19), where two terms are balanced by a choice of ‘learning rate’ comes up very often in online learning [93], which hints at a connection between K-learning and these techniques.

Algorithm 1 K-learning for multi-armed bandits

Input: multi-armed bandit $\{A, R\}$, uncertainty parameter σ
initialize $n_a \leftarrow 0$, $a = 1, \dots, A$, $\mathcal{F}_1 \leftarrow \emptyset$
for round $t = 1, 2, \dots$, **do**
 calculate $\beta_t = \sqrt{\frac{4t \log A}{\sigma^2 A(1+\log t)}}$ or $\beta_t = \operatorname{argmin}_{\beta} ((1/\beta) \log \sum \exp K_{\mu_i}^t(\beta))$
 calculate posterior mean of rewards $\hat{\mu} = \mathbf{E}(\mu|\mathcal{F}_t)$
 calculate k values for $a = 1, \dots, A$: $k_a = \hat{\mu}_a + \frac{\sigma^2 \beta_t}{2(1+n_a)}$
 sample action a_t with probability $\pi_a \propto \exp(\beta_t k_a)$
 update $n_{a_t} \leftarrow n_{a_t} + 1$, $\mathcal{F}_{t+1} \leftarrow \mathcal{F}_t \cup \{a_t, r_t\}$
end for

3.1 Gap-dependent bound

Many other multi-armed bandit algorithms have an instance bound that depends on the gap between the best and second best arm and grows logarithmically with time, rather than as the square root. For example, in UCB the regret for any instance is bounded by $O(\log T/\Delta)$, where Δ is the gap. One might ask if the same logarithmic bound holds for K-learning. In this section we briefly sketch a negative result that such a bound is unlikely to hold for the fixed β_t schedule, though this doesn't say anything about the regret when using the optimal β_t as determined by minimizing the quantity in eq. (16). Consider a case with two arms, where the gap is Δ . The regret is given by the number of times the sub-optimal arm, denoted N , is pulled multiplied by Δ , *i.e.*, $R(T) \approx N\Delta$. The sub-optimal arm is likely to get selected when the bonus of the arm plus the expected value is close to the optimal expected value, *i.e.*, consider when

$$\mu^* + \sqrt{T}/(T - N) \approx \mu^* - \Delta + \sqrt{T}/N$$

if the first arm is pulled more often then $\sqrt{T}/(T - N)$ is small, and so the regret $R(T) \approx N\Delta \approx \sqrt{T}$.

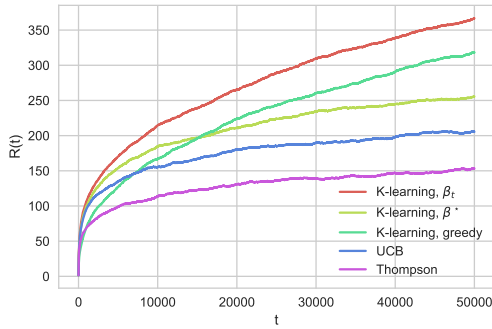
Intuitively speaking the bonus afforded to the arms in K-learning is 'too high' to permit a regret bound that grows logarithmically. In other words, the algorithm is 'too exploratory'. However, this extra exploration may be useful in the case of adversarial bandits, or where the Bayesian prior is misspecified. We shall show that this is the case in the experiments.

3.2 Numerical experiments

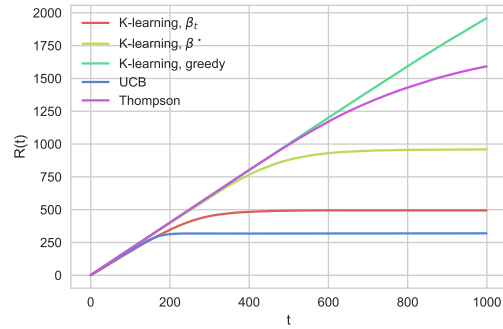
We compare the performance of several bandit algorithms in Fig. 2 under different conditions. In all three experiments we ran K-learning with the fixed schedule for β_t , K-learning with the optimal β_t , greedy K-learning (described in the appendix), Thompson sampling, and UCB, and we report regret averaged over 500 runs. Fig. 2a compares the algorithms under 'nominal' conditions with 10 arms, where the prior for the mean reward of each arm is $\mathcal{N}(0, 1)$, and the reward noise was $\mathcal{N}(0, 1)$. In this case Thompson sampling performs best, UCB is next and K-learning performs worst. In Fig. 2b we compare the same algorithms, except in this case we have misspecified the prior. In theory the Bayesian regret is bounded by a norm of the Radon-Nikodym derivative of the true prior with respect to the assumed prior [87, §3.1], however in practice different algorithms may demonstrate different levels of robustness. Specifically, in this example there are just two arms, where the prior for the first arm is $\mathcal{N}(-1, 0.1)$ and the second arm is $\mathcal{N}(1, 0.1)$, however when running the algorithm we swap the priors, so the mean reward of the first arm is sampled from the second prior and vice-versa. In this case UCB performs well, since it does not use prior information at

all. However Thompson sampling performs very poorly, and K-learning performs almost as well as UCB. In Fig. 2c we compare what happens under ‘pseudo-adversarial’ conditions. We consider two arms, where initially the first arm gives a reward of 1 and the second arm gives a reward of 0, then just before the half-way mark, the rewards swap and the second arm becomes the better arm. Since the arms swapped performance before the half-way mark it is the second arm that is the better arm overall and how quickly the algorithms can discover this fact will determine their overall performance (this is not truly the adversarial setting, and we make no claim on the regret of K-learning under true adversarial conditions [7]). As we see K-learning is by far the quickest to respond and ultimately ends with *negative* regret, Thompson sampling also adjusts, albeit significantly more slowly. UCB by comparison does not adjust in time and ends with positive regret. Note that under real adversarial conditions any greedy algorithm, like UCB or greedy K-learning, would be easily exploited. In figure 3 we plot the performance of K-learning using the β schedule and K-learning using the optimal choice of β and the associated upper bound as determined by (15).

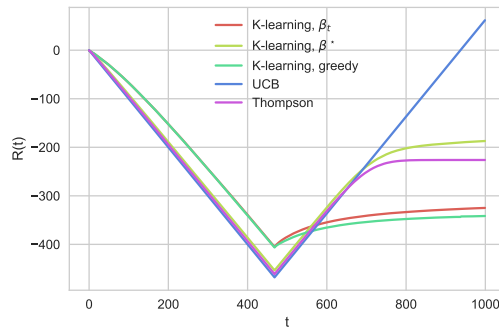
In conclusion, under nominal conditions Thompson sampling and UCB both outperform K-learning. However, if the priors are misspecified or the environment is changing or adversarial, a realistic proposition in practice, then K-learning performs much better. In other words, due to the fact that K-learning explores more it generally has worse regret under totally nominal conditions, however, empirically speaking, it is more robust to changing conditions and model errors.



(a) Nominal conditions.



(b) Misspecified prior.



(c) Pseudo-adversarial.

Figure 2: Comparison of bandit algorithms under different conditions.

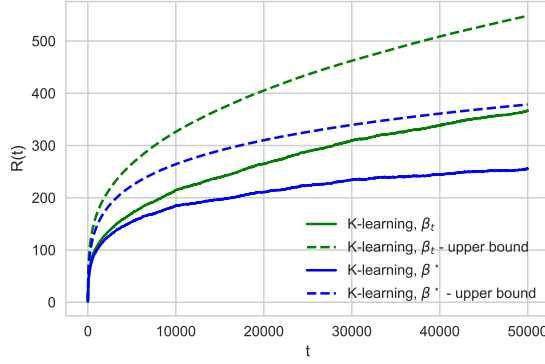


Figure 3: Regret and expected regret upper bound.

4 Markov decision processes

In a Markov decision process (MDP) an agent interacts with an environment in a series of episodes and attempts to maximize its long-term return. A finite horizon discrete MDP is given by the tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, R, P, L, \rho\}$, where $\mathcal{S} = \{1, \dots, S\}$ is the state-space, $\mathcal{A} = \{1, \dots, A\}$ is the action-space, $R(s, a)$ is a probability distribution over the rewards received by the agent at state s taking action a , and $P(s'|s, a)$ is a probability the agent will transition to state s' after taking action a in state s , $L \in \mathbb{N}$ is the episode length, and ρ is the initial state distribution. Concretely, the initial state of the agent is sampled from ρ , then for time-steps $l = 1, \dots, L$ the agent is in state $s_l \in \mathcal{S}$, selects action $a_l \in \mathcal{A}$, receives reward $r_l \sim R(s_l, a_l)$ and transitions to the next state $s_{l+1} \sim P(\cdot|s_l, a_l)$. After time-step L the episode terminates and the state is reset. We assume that at the beginning of learning the agent does not know the reward or transition probabilities and must learn about them by interacting with the environment, and we consider the Bayesian case, where the mean rewards and the transition probabilities are sampled from known priors. In the previous section we dealt with the stochastic multi-armed bandit, which is a special case of an MDP with $S = L = 1$. For the analysis we will make the following assumption:

Assumption 2. *The MDP is a directed acyclic graph (DAG).*

This assumption implies that a state cannot be visited again within the same episode. This may seem like a strong assumption, but in fact any finite horizon MDP can be converted into one that satisfies this assumption by ‘unrolling’ the MDP over time, whereby each state is replaced by L copies of itself and each copy indexed with the corresponding time-step l for $l = 1, \dots, L$. If we do this then in the worst case the state space cardinality goes from S to LS , which will affect the regret bound we derive. So with this in mind, hereafter we will use \mathcal{X} to represent the state-space, where

$$|\mathcal{X}| = \begin{cases} S & \text{if assumption 2 holds} \\ LS & \text{otherwise.} \end{cases}$$

This assumption allows us to take the state space to be the union of L distinct sets, *i.e.*, $\mathcal{X} = \mathcal{S}_1 \cup \dots \cup \mathcal{S}_L$, where at time-step l the agent is in state $s_l \in \mathcal{S}_l$, and so $|\mathcal{X}| = \sum_{l=1}^L |\mathcal{S}_l|$.

An agent following policy π selects actions as $a_l \sim \pi(s_l)$. The Bellman equation relates the value of actions taken at the current time-step to future returns through the Q -values [15], which for policy π are

given by

$$Q^\pi(s_l, a_l) = \mu(s_l, a_l) + \sum_{s_{l+1}} P(s_{l+1}|s_l, a_l) \sum_{a_{l+1}} \pi(s_{l+1}, a_{l+1}) Q^\pi(s_{l+1}, a_{l+1})$$

for $l = 1, \dots, L$ where $Q(s_{L+1}, \cdot)$ is defined to be zero, $\mu = \mathbf{E} r$ is the mean reward, the next state s_{l+1} is drawn from $P(\cdot|s_l, a_l)$, and the next action a_{l+1} is drawn from $\pi(s_{l+1})$. The value function captures the expected value of a particular state under a policy π

$$V^\pi(s_l) = \sum_{a_l} \pi(s_l, a_l) Q^\pi(s_l, a_l).$$

The performance of a policy is given by $J^\pi = \mathbf{E}_{s_1 \sim \rho} V^\pi(s_1)$. An optimal policy satisfies $\pi^* \in \operatorname{argmax}_\pi J^\pi$ and induces associated Q-values given by the optimal Bellman equation

$$Q^*(s_l, a_l) = \mu(s_l, a_l) + \sum_{s_{l+1}} P(s_{l+1}|s_l, a_l) \max_{a_{l+1}} Q^*(s_{l+1}, a_{l+1}) \quad (20)$$

for $l = 1, \dots, L$, with associated value functions

$$V^*(s_l) = \max_{a_l} Q^*(s_l, a_l).$$

If the rewards and transition function are known exactly then we could solve (20) via dynamic programming [16]. However, in practice these are not known and so the agent must gather data by interacting with the environment. The key trade-off is the *exploration-exploitation* dilemma, whereby an agent must trade-off exploring to learn more about the environment and taking rewards it has less uncertainty about. As we shall see, imbuing the agent with an epistemic risk-seeking utility provides a principled approach to this dilemma.

In this section we extend the regret analysis for K-learning applied to multi-armed bandits of the previous section to the episodic MDP case. We shall index the episode number with t , and elapsed time-steps in the current episode as l . The Bayesian expected regret of any policy after T time-steps is defined to be

$$\mathbf{E} R(T) = \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} \sum_{s_1} \rho(s_1) (V^*(s_1) - V^{\pi_t}(s_1))$$

where the outer expectation is with respect to the priors over the reward and state transition functions (the stochasticity of the rewards and policy is already accounted for in the definition of V^* and V^π). Since the reward and transition functions are random variables (drawn from the priors) it implies that the Q-values are also random variables, and we can consider the posterior over the Q-values and use them to generate a policy. We shall make use of the following inequality

$$\mathbf{E}(\max_a Q(s, a) | \mathcal{F}_t) \leq (1/\beta) \log \sum_a \exp K_{Q^*}^t(s, a, \beta),$$

where as before \mathcal{F}_t denotes the sigma-algebra generated by all the history before time t and where $K_{Q^*}^t(s, a, \cdot)$ denotes the cumulant generating function of the posterior over the Q-value at state-action s, a conditioned on \mathcal{F}_t . Similarly we will denote by $K_\mu^t(s, a, \cdot)$ the cumulant generating function of the posterior of the *mean* reward at state-action s, a and conditioned on \mathcal{F}_t . For our analysis we will require an assumption that is the analogue of assumption 1 which we made for bandits:

Assumption 3. *The posterior over the mean value of the reward at each state-action is sub-Gaussian, and concentrates with data at least as fast as a Gaussian, i.e.,*

$$\mathbf{E}(\exp \beta(\mu(s, a) - \mathbf{E}(\mu(s, a)|\mathcal{F}_t))|\mathcal{F}_t) \leq \exp(\beta^2 \sigma^2 / 2n_t(s, a))$$

for some $\sigma \geq 0$ and where $n_t(s, a)$ is the number of times we have been in state s and taken action a after t time-steps.

4.1 The case with known dynamics

Before we consider solving the MDP problem in full generality we start with the simpler situation where the reward function R is unknown but the transition probabilities P are known. We shall prove a regret bound for this case which depends only on the uncertainty in the mean returns. In the sequel we will incorporate uncertainty about the transition probabilities as well as the rewards, and the analysis we perform here will be useful. We begin by proving a lemma on the cumulant generating functions of the posterior distributions. In a slight abuse of notation we shall use $\mathbf{E}_{s_{l+1}}$ to denote the expectation operator under P , i.e., $\mathbf{E}_{s_{l+1}} x(s_{l+1}) := \sum_{s_{l+1}} P(s_{l+1}|s_l, a_l) x(s_{l+1})$ for any $x \in \mathbb{R}^{|S_{l+1}|}$.

Lemma 6. *When the transition function is known the cumulant generating function for the posterior over the Q -values satisfies a Bellman inequality*

$$K_{Q^*}^t(s_l, a_l, \beta) \leq K_\mu^t(s_l, a_l, \beta) + \mathbf{E}_{s_{l+1}} \log \sum_{a_{l+1}} \exp K_{Q^*}^t(s_{l+1}, a_{l+1}, \beta). \quad (21)$$

Proof. This follows directly from the definition of the cumulant generating function

$$\begin{aligned} K_{Q^*}^t(s_l, a_l, \beta) &= \log \mathbf{E}(\exp \beta Q^*(s_l, a_l) | \mathcal{F}_t) \\ &\stackrel{(a)}{=} \log \mathbf{E}(\exp \beta(\mu(s_l, a_l) + \mathbf{E}_{s_{l+1}} \max_{a_{l+1}} Q^*(s_{l+1}, a_{l+1})) | \mathcal{F}_t) \\ &\stackrel{(b)}{=} K_\mu^t(s_l, a_l, \beta) + \log \mathbf{E}(\exp \mathbf{E}_{s_{l+1}} \max_{a_{l+1}} \beta Q^*(s_{l+1}, a_{l+1}) | \mathcal{F}_t) \\ &\stackrel{(c)}{\leq} K_\mu^t(s_l, a_l, \beta) + \mathbf{E}_{s_{l+1}} \log \mathbf{E}(\exp \max_{a_{l+1}} \beta Q^*(s_{l+1}, a_{l+1}) | \mathcal{F}_t) \\ &\stackrel{(d)}{\leq} K_\mu^t(s_l, a_l, \beta) + \mathbf{E}_{s_{l+1}} \log \sum_{a_{l+1}} \exp K_{Q^*}^t(s_{l+1}, a_{l+1}, \beta), \end{aligned}$$

where (a) is the Bellman equation (20), (b) follows because the MDP is a DAG by assumption 2, (c) is due to Jensen's inequality, and (d) is that the max of a collection of nonnegative numbers is less than the sum. \square

Corollary 4. *The certainty equivalent value under the exponential risk-seeking utility satisfies a Bellman inequality*

$$C_{Q^*}(s_l, a_l, \beta) \leq (1/\beta) K_\mu^t(s_l, a_l, \beta) + \mathbf{E}_{s_{l+1}} (1/\beta) \log \sum_{a_{l+1}} \exp \beta C_{Q^*}(s_{l+1}, a_{l+1}, \beta).$$

Armed with this lemma we can analyze the regret of the Boltzmann policy over the cumulant generating functions.

Lemma 7. *In the case where the transition function P is known, the variational policy*

$$\pi_t(s, a, \beta_t) \propto \exp K_{Q^*}^t(s, a, \beta_t), \quad (22)$$

where the inverse temperature parameter β_t is given by

$$\beta_t = \sqrt{\frac{4tL \log A}{\sigma_r^2 A |\mathcal{X}| (1 + \log t)}} \quad (23)$$

which is fixed for the entirety of each episode t , has Bayesian expected regret bounded by

$$\begin{aligned} \mathbf{E} R(T) &\leq 2\sigma_r \sqrt{T|\mathcal{X}|A \log A(1 + \log T/L)} \\ &= \begin{cases} \tilde{O}(\sqrt{TSA}) & \text{if assumption 2 holds} \\ \tilde{O}(\sqrt{L TSA}) & \text{otherwise.} \end{cases} \end{aligned}$$

Proof. Again we use the tower property of conditional expectation together with lemma 3 and the fact that π_t is \mathcal{F}_t -measurable

$$\begin{aligned} \mathbf{E} R(T) &= \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} \mathbf{E}_{s_1} (V^*(s_1) - V^{\pi_t}(s_1)) \\ &= \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} \mathbf{E}_{s_1} \mathbf{E}(\max_a Q^*(s_1, a) - \sum_{a_1} \pi_t(s_1, a_1) Q^{\pi_t}(s_1, a_1) | \mathcal{F}_t) \\ &\leq \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} \mathbf{E}_{s_1} (1/\beta_t) \left(\sum_{a_1} \pi_t(s_1, a_1) (K_{Q^*}^t(s_1, a_1, \beta_t) - \beta_t \mathbf{E}(Q^{\pi_t}(s_1, a_1) | \mathcal{F}_t)) + H(\pi_t(s_1)) \right) \end{aligned}$$

Now we recursively apply the following inequality

$$\begin{aligned} K_{Q^*}^t(s_l, a_l, \beta_t) - \beta_t \mathbf{E}(Q^{\pi_t}(s_l, a_l) | \mathcal{F}_t) &\leq K_{\mu - \mathbf{E}(\mu | \mathcal{F}_t)}^t(s_l, a_l, \beta_t) + \mathbf{E}_{s_{l+1}} H(\pi_t(s_{l+1})) + \\ &\quad \mathbf{E}_{s_{l+1}} \sum_{a_{l+1}} \pi_t(s_{l+1}, a_{l+1}) (K_{Q^*}^t(s_{l+1}, a_{l+1}, \beta_t) - \beta_t \mathbf{E}(Q^{\pi_t}(s_{l+1}, a_{l+1}) | \mathcal{F}_t)) \end{aligned}$$

for $l = 1, \dots, L$, which comes from expanding the Q-value using the Bellman equation and upper bounding the cumulant generating function using lemmas 1 and 6, as well as using the fact that π_t is \mathcal{F}_t -measurable. This implies we accumulate L terms, each of which is the same as we had for the bandit case in eq. 14. Let $\lambda(s_l, a_l)$ be the occupancy measure for state s_l and action a_l , i.e., the probability of being in state s_l and taking action a_l under the policy π , then we have

$$\begin{aligned} \mathbf{E} R(T) &\leq \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} (1/\beta_t) \sum_{l=1}^L \sum_{s_l, a_l} \lambda(s_l, a_l) (H(\pi(s_l)) + K_{\mu - \mathbf{E}(\mu | \mathcal{F}_t)}^t(s_l, a_l, \beta_t)) \\ &\stackrel{(a)}{\leq} \mathbf{E} \sum_{t=1}^{\lceil T/L \rceil} (1/\beta_t) (L \log A + \sum_{l=1}^L \sum_{s_l, a_l} \lambda(s_l, a_l) \sigma_r^2 \beta_t^2 / n_t(s_l, a_l)) \\ &\stackrel{(b)}{\leq} \sum_{t=1}^{\lceil T/L \rceil} (L \log A / \beta_t) + (\sum_{l=1}^L |S_l|) A \sigma_r^2 \beta_{\lceil T/L \rceil} (1 + \log T/L) \\ &\stackrel{(c)}{\leq} 2\sigma_r \sqrt{T|\mathcal{X}|A \log A(1 + \log T/L)}, \end{aligned} \quad (24)$$

in (a) we use the bound on entropy and the sub-Gaussian assumption 3 and for (b) we applied lemma 5 to each of the L terms, and for (c) we substitute β_t from eq. 23. \square

5 K-learning

In the previous section on MDPs we assumed knowledge of the cumulant generating function for the posterior over the Q-values. In practice this is generally intractable to compute. In this section we introduce K-learning, an extension of Q-learning where instead of the true cumulant generating function we use an approximation, which can be interpreted as an upper bound on the certainty equivalent value for the risk-seeking utility. We start again with the simpler case, that of known dynamics.

5.1 The case with known dynamics

Consider the K-learning Bellman equation given by

$$\mathcal{K}(s_l, a_l, \beta) = K_\mu(s_l, a_l, \beta) + \mathbf{E}_{s_{l+1}} \log \sum_{a_{l+1}} \exp \mathcal{K}(s_{l+1}, a_{l+1}, \beta) \quad (25)$$

for all $\beta \geq 0$, and for $l = 1, \dots, L$, where $\mathcal{K}(s_{L+1}, \cdot, \cdot) \equiv 0$.

Lemma 8. *In the case with known transition function, the K-function \mathcal{K} is a pointwise upper bound on the true cumulant generating function K .*

Proof. For shorthand let $K_{Q_l^*} = K_{Q^*}|_{S_l}$ and $\mathcal{K}_l = \mathcal{K}|_{S_l}$, i.e., the functions K_{Q^*} and \mathcal{K} restricted to states in S_l , and write the recursion in eq. (25) for some fixed $\beta \geq 0$ in operator notation as $\mathcal{K}_l = \mathcal{T}_l \mathcal{K}_{l+1}$. Since log-sum-exp is increasing it implies that the operator \mathcal{T}_l is monotonically increasing in its argument, i.e., if $x \geq y$ then $\mathcal{T}_l x \geq \mathcal{T}_l y$ pointwise. From equation (21) we have that $K_{Q_{l+1}^*} \leq \mathcal{T}_l K_{Q_l^*}$ for each l . The proof follows by induction; assume $\mathcal{K}_{l+1} \geq K_{Q_{l+1}^*}$, then

$$\mathcal{K}_l = \mathcal{T}_l \mathcal{K}_{l+1} \geq \mathcal{T}_l K_{Q_{l+1}^*} \geq K_{Q_l^*},$$

and the base case holds since $\mathcal{K}_{L+1} \equiv K_{Q_{L+1}^*} \equiv 0$. \square

Lemma 9. *In the case of known state-transition function, replacing the true cumulant generating function in (22) with the K-function \mathcal{K} achieves the same regret bound, given by (24).*

Proof. Lemma (8) implies that

$$\mathbf{E}(\max_a Q^*(s, a) | \mathcal{F}_t) \leq (1/\beta) \log \sum_a \exp \mathcal{K}^t(s, a, \beta),$$

since log-sum-exp is increasing. To conclude the regret bound we note that the only other property we required was as in equation (21), which \mathcal{K} satisfies by construction. Since we follow the policy as induced by \mathcal{K} then the proof follows immediately. \square

Lemma 9 only yields a regret bounds when the transition function is known. Next we extend the regret bound to the case where the transition function is unknown.

5.2 The case with unknown dynamics

In the analysis up to now we have assumed that the dynamics are stochastic, but the state-transition function is known. In this section we remove that assumption and prove a regret bound for the full K-learning algorithm. In order to prove a regret bound here we make two additional assumptions.

Assumption 4. *The mean rewards are bounded in $[0, 1]$.*

This only affects the *mean* reward, the noise in the rewards may be outside this range, in which case we would still require the sub-Gaussian assumption 3. If the rewards themselves are bounded in $[0, 1]$ then this implies that the rewards are sub-Gaussian with $\sigma_r = 1/2$, due to Hoeffding's lemma. Our analysis is easily extended to the case where the rewards are bounded in some $[R_{\min}, R_{\max}]$. We need this assumption in order to bound the span of the optimal value function, that is $\text{Span}(V^*) = \max_s V^*(s) - \min_s V^*(s)$,

which we will require later. We could replace this assumption with an assumption on knowledge of a bound on the span, but an assumption on bounded rewards is more common in the literature.

Now that we are considering unknown transition functions we make an additional assumption on the form of the prior over the transition functions P .

Assumption 5. *The prior for $P(\cdot|s_l, a_l)$ is Dirichlet with parameter $\alpha_l^0(\cdot, s_l, a_l) \in \mathbb{R}_+^{|\mathcal{S}_{l+1}|}$, for each $(s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}$, $l = 1, \dots, L$.*

Since the likelihood for the transition function is a Categorical distribution, conjugacy of the categorical and Dirichlet distributions implies that the posterior over $P(s_{l+1}|s_l, a_l)$ at time t is Dirichlet with parameter $\alpha_l^t(s_{l+1}, s_l, a_l)$, where

$$\alpha_l^t(s_{l+1}, s_l, a_l) = \alpha_l^0(s_{l+1}, s_l, a_l) + n_t(s_{l+1}|s_l, a_l)$$

for each $s_{l+1} \in \mathcal{S}_{l+1}$, where $n_t(s_{l+1}|s_l, a_l) \in \mathbb{N}$ is the count of times the agent has been in state s_l , taken action a_l , and transitioned to state s_{l+1} , and note that $\sum_{s_{l+1} \in \mathcal{S}_{l+1}} n_t(s_{l+1}|s_l, a_l) = n_t(s_l, a_l)$, the total visit count to (s_l, a_l) . Our analysis will make use of the following definition and associated lemma from Osband and Van Roy [76]:

Definition 1. *Let X and Y be random variables, we say that X is stochastically optimistic for Y , written $X \geq_{SO} Y$, if $\mathbf{E} u(X) \geq \mathbf{E} u(Y)$ for any convex increasing function u .*

Stochastic optimism is closely related to the more familiar concept of second-order stochastic dominance, in that X is stochastically optimistic for Y if and only if $-Y$ second-order stochastically dominates $-X$ [32]. We use this definition in the next lemma.

Lemma 10. *Let $Y = \sum_{i=1}^n A_i b_i$ for fixed $b \in \mathbb{R}^n$ and random variable A , where A is Dirichlet with parameter $\alpha \in \mathbb{R}^n$, and let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ with $\mu_X \geq \frac{\sum_i \alpha_i b_i}{\sum_i \alpha_i}$ and $\sigma_X^2 \geq (\sum_i \alpha_i)^{-1} \text{Span}(b)^2$, where $\text{Span}(b) = \max_i b_i - \min_j b_j$, then $X \geq_{SO} Y$.*

We will use these to derive a Bellman inequality analogous to lemma 9 for the case where the transition function is unknown. For shorthand we shall use the notation $\hat{\mathbf{E}}_{s_{l+1}}$ to denote the expectation operator induced by the mean of the posterior over state transition function P conditioned on \mathcal{F}_t , i.e., $\hat{\mathbf{E}}_{s_{l+1}} x(s_{l+1}) := \sum_{s_{l+1}} \mathbf{E}(P(s_{l+1}|s_l, a_l)|\mathcal{F}_t) x(s_{l+1})$ for any $x \in \mathbb{R}^{|\mathcal{S}_{l+1}|}$.

Lemma 11. *Under above assumptions the cumulant generating function for the posterior over the Q -values satisfy a Bellman inequality*

$$K_{Q^*}^t(s_l, a_l, \beta) \leq K_\mu^t(s_l, a_l, \beta) + L^2 \beta^2 / n_t(s_l, a_l) + \hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp K_{Q^*}^t(s_{l+1}, a_{l+1}, \beta). \quad (26)$$

Proof. In our case, in the notation of the lemma 10, A will represent the transition function probabilities, and b will represent the optimal values of the next state, i.e., for a given $(s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}$ we let $Y := \sum_{s_{l+1}} P(s_{l+1}|s_l, a_l) V^*(s_{l+1})$ and X_t be a random variable distributed $\mathcal{N}(\mu_{X_t}, \sigma_{X_t}^2)$ where

$$\begin{aligned} \mu_{X_t} &= \sum_{s_{l+1} \in \mathcal{S}_{l+1}} \left(\alpha_l^t(s_{l+1}, s_l, a_l) V^*(s_{l+1}) / \sum_{x \in \mathcal{S}_{l+1}} \alpha_l^t(x, s_l, a_l) \right) \\ &\stackrel{(a)}{=} \sum_{s_{l+1} \in \mathcal{S}_{l+1}} \mathbf{E}(P(s_{l+1}|s_l, a_l)|\mathcal{F}_t) V^*(s_{l+1}) \\ &= \hat{\mathbf{E}}_{s_{l+1}} V^*(s_{l+1}) \end{aligned}$$

where (a) is due to the Dirichlet assumption 5. Due to assumption 4 we know that $\text{Span}(V^*) \leq L$, so we choose σ_{X_t} to satisfy $\sigma_{X_t}^2 \leq L^2/n_t(s_l, a_l)$. Let $\mathcal{F}_t^V = \mathcal{F}_t \cup \sigma(V^*)$ denote the union of \mathcal{F}_t and the sigma-algebra generated by V^* . Applying lemma 10 and the tower property of conditional expectation we have that for $\beta \geq 0$

$$\mathbf{E}_{P, V^*}(\exp \beta Y | \mathcal{F}_t) = \mathbf{E}_{V^*}(\mathbf{E}_P(\exp \beta Y | \mathcal{F}_t^V) | \mathcal{F}_t) \leq \mathbf{E}_{V^*}(\mathbf{E}_{X_t}(\exp \beta X_t | \mathcal{F}_t^V) | \mathcal{F}_t) = \mathbf{E}_{V^*}(\exp(\mu_X \beta + \sigma_X^2 \beta^2) | \mathcal{F}_t), \quad (27)$$

where the last equality comes from the moment-generating function of a Gaussian. Putting it all together

$$\begin{aligned} \log \mathbf{E}_{P, V^*} \exp(\beta \sum_{s_{l+1}} P(s_{l+1} | s_l, a_l) V^*(s_{l+1}) | \mathcal{F}_t) \\ &\stackrel{(a)}{\leq} \log \mathbf{E}_{V^*}(\exp(\mu_X \beta + \sigma_X^2 \beta^2) | \mathcal{F}_t) \\ &\stackrel{(b)}{=} \log \mathbf{E}_{Q^*}(\exp(\beta \hat{\mathbf{E}}_{s_{l+1}} \max_{a_{l+1}} Q^*(s_{l+1}, a_{l+1}) | \mathcal{F}_t) + L^2 \beta^2 / n_t(s_l, a_l) \quad (28) \\ &\stackrel{(c)}{\leq} \hat{\mathbf{E}}_{s_{l+1}} \log \mathbf{E}_{Q^*}(\exp \beta \max_{a_{l+1}} Q^*(s_{l+1}, a_{l+1}) | \mathcal{F}_t) + L^2 \beta^2 / n_t(s_l, a_l) \\ &\stackrel{(d)}{\leq} \hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp K_{Q^*}^t(s_{l+1}, a_{l+1}, \beta) + L^2 \beta^2 / n_t(s_l, a_l) \end{aligned}$$

where (a) follows from eq. (27) and the fact that \log is increasing, (b) is substituting out μ_{X_t} and σ_{X_t} and replacing V^* with Q^* , (c) is Jensen's inequality, and (d) follows since the max of a collection of positive numbers is less than the sum. From this and using the logic in lemma 6 the inequality immediately follows. \square

Armed with lemma 11 we can define an updated Bellman equation for the K-functions when the transition function is unknown

$$\mathcal{K}^t(s_l, a_l, \beta) = K_\mu^t(s_l, a_l, \beta) + L^2 \beta^2 / n_t(s_l, a_l) + \hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp \mathcal{K}_{l+1}^t(s_{l+1}, a_{l+1}, \beta). \quad (29)$$

Given this equation we can formalize our method into an algorithm, presented as algorithm 2.

Theorem 2. *The K-learning algorithm 2 achieves Bayesian expected regret bound*

$$\begin{aligned} \mathbf{E} R(T) &\leq 2\sqrt{(\sigma_r^2 + L^2)T|\mathcal{X}|A \log A(1 + \log T/L)} \\ &= \begin{cases} \tilde{O}(L\sqrt{TSA}) & \text{if assumption 2 holds} \\ \tilde{O}(L^{3/2}\sqrt{TSA}) & \text{otherwise.} \end{cases} \end{aligned}$$

Proof. Consider equation (29) and observe that we could rewrite it as

$$\mathcal{K}^t(s_l, a_l, \beta) = \tilde{K}_\mu^t(s_l, a_l, \beta) + \hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp \mathcal{K}_{l+1}^t(s_{l+1}, a_{l+1}, \beta)$$

where $\tilde{K}_\mu^t = K_\mu^t(s_l, a_l, \beta) + L^2 \beta^2 / n_t(s, a)$ and which, due to the sub-Gaussian assumption on K_μ , now satisfies

$$\tilde{K}_{\mu - \mathbf{E}(\mu | \mathcal{F}_t)}^t(s_l, a_l, \beta) \leq (\sigma_r^2 + L^2) \beta^2 / 2n_t(s_l, a_l).$$

In other words we can replace the MDP with unknown rewards and transitions with one where only the rewards are unknown but with higher uncertainty, and where the transition function is taken to be the mean

of the posterior at each episode. This fact allows us to apply the regret bound we already had in the previous section, the only additional fact we need is the following

$$\begin{aligned}
\mathbf{E}(Q^\pi(s_l, a_l) | \mathcal{F}_t) &\stackrel{(a)}{=} \mathbf{E}(r(s_l, a_l) + \sum_{s_{l+1} \in \mathcal{S}_{l+1}} P(s_{l+1} | s_l, a_l) \sum_{a_{l+1} \in \mathcal{A}} \pi(s_{l+1}, a_{l+1}) Q^\pi(s_{l+1}, a_{l+1}) | \mathcal{F}_t) \\
&\stackrel{(b)}{=} \mathbf{E}(r(s_l, a_l) | \mathcal{F}_t) + \sum_{s_{l+1} \in \mathcal{S}_{l+1}} \mathbf{E}(P(s_{l+1} | s_l, a_l) | \mathcal{F}_t) \sum_{a_{l+1} \in \mathcal{A}} \mathbf{E}(\pi(s_{l+1}, a_{l+1}) Q^\pi(s_{l+1}, a_{l+1}) | \mathcal{F}_t) \\
&\stackrel{(c)}{=} \mathbf{E}(\mu(s_l, a_l) | \mathcal{F}_t) + \hat{\mathbf{E}}_{s_{l+1}} \sum_{a_{l+1} \in \mathcal{A}} \pi(s_{l+1}, a_{l+1}) \mathbf{E}(Q^\pi(s_{l+1}, a_{l+1}) | \mathcal{F}_t)
\end{aligned}$$

where (a) is the Bellman equation (20), (b) holds due to the fact that the MDP is a DAG from assumption 2, (c) holds since π_t is \mathcal{F}_t measurable and using eq. (13). With these facts we can replicate the procedure of lemma 9 to obtain the regret bound, setting β_t to

$$\beta_t = \sqrt{\frac{4tL \log A}{(\sigma_r^2 + L^2)A|\mathcal{X}|(1 + \log t)}}. \quad (30)$$

Finally, note that since we know the value of β_t for any episode we don't need to compute the full K-function, only the K-functions at the value of β_t we are interested in. So if we define $k(s, a) = \mathcal{K}(s, a, \beta_t) / \beta_t$ then we can rewrite equation (29) using this K-value, which yields eq. (31). Note that

$$\mathbf{E}(Q^*(s, a) | \mathcal{F}_t) \leq C_{Q^*}(s, a, \beta_t) = (1/\beta_t) K_{Q^*}(s, a, \beta_t) \leq (1/\beta_t) \mathcal{K}(s, a, \beta_t) = k(s, a),$$

so in other words the K-values are *optimisitic* for both the expected Q-values under the posterior and the certainty equivalent value of the exponential utility function. \square

The Bayesian expected regret bound in the above theorem is within a factor of L of the known minimax lower bound of

$$R(T) \geq \Omega(\sqrt{LSAT}),$$

though this is not necessarily a lower bound on the expected regret for the restricted class of problems we consider due to the assumptions we made on the prior of P .

5.2.1 Greedy formulation

For any collection of numbers $X_i, i = 1, \dots, A$, we have $\log \sum_{i=1}^k \exp X_i \leq \max_i X_i + \log A$, and so we can replace the K-learning update (31) with

$$k(s_l, a_l) = \hat{\mu}(s_l, a_l) + \frac{(\sigma_r^2 + L^2)\beta_t}{2(1 + n_t(s_l, a_l))} + \hat{\mathbf{E}}_{s_{l+1}} \max_{a_{l+1}} k_{l+1}(s_{l+1}, a_{l+1}) + \log A / \beta_t$$

and maintain the upper bound property. These values combined with the greedy policy whereby at state s we select any $a \in \arg\max_b k(s, b)$ has the same regret bound as K-learning and it is a straightforward change to the proof. Since the $\log A / \beta_t$ term affects all K-values at each time-step equally we can remove it and not change the policy. The full greedy algorithm is detailed in the appendix.

Algorithm 2 K-learning for episodic MDPs

Input: MDP $\mathcal{M} = \{\mathcal{X}, \mathcal{A}, R, P, L, \rho\}$,
initialize $n(s, a) \leftarrow 0$ for all $s, a \in \mathcal{X} \times \mathcal{A}$, $\mathcal{F}_1 = \emptyset$
for episode $t = 1, 2, \dots$ **do**
 calculate $\beta_t = \sqrt{\frac{4tL \log A}{(\sigma_r^2 + L^2)A|\mathcal{X}|(1+\log t)}}$
 calculate $\hat{\mu} = \mathbf{E}(\mu|\mathcal{F}_t)$ and transition operator $\hat{\mathbf{E}}_{s_{l+1}}$ via $\mathbf{E}(P|\mathcal{F}_t)$
 let $k(s_{L+1}, \cdot) = 0$
 for step $l = L, \dots, 1$ **do**

$$k(s_l, a_l) = \hat{\mu}(s_l, a_l) + \frac{(\sigma_r^2 + L^2)\beta_t}{2(1 + n_t(s_l, a_l))} + (1/\beta_t)\hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp \beta_t k(s_{l+1}, a_{l+1}) \quad (31)$$

for each $(s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}$
 end for
 for step $l = 1, \dots, L$ **do**
 at state s_l sample action a_l with probability $\pi(a_l|s_l) \propto \exp(\beta_t k(s_l, a_l))$
 end for
 update $n(s, a) \leftarrow n(s, a) + 1$ for visited s, a
 update $\mathcal{F}_{t+1} \leftarrow \mathcal{F}_t \cup \{s_l, a_l, r_l, s_{l+1} : l = 1, \dots, L\}$
end for

5.3 Formulation as a convex optimization problem

Analogous to the bandit case, we can formulate the problem of finding the optimal choice of $\beta = 1/\tau$ at each episode as a convex optimization problem. We want to find the choice of τ that provides the tightest upper bound in the family for $\mathbf{E}_{s_1 \sim \rho} \mathbf{E}(\max_a Q^*(s_1, a)|\mathcal{F}_t)$, which can be represented as the following convex optimization problem

$$\begin{aligned} & \text{minimize} && \mathbf{E}_{s_1 \sim \rho} \tau \log \sum_{a_1} \exp k(s_1, a_1)/\tau \\ & \text{subject to} && k(s_l, a_l) \geq \tau \tilde{K}_\mu^t(s_l, a_l, 1/\tau) + \tau \hat{\mathbf{E}}_{s_{l+1}} \log \sum_{a_{l+1}} \exp k(s_{l+1}, a_{l+1})/\tau, \\ & && s_l \in \mathcal{S}_l, \quad a_l \in \mathcal{A}, \quad l = 1, \dots, L \\ & && k(s_{L+1}, \cdot) = 0 \end{aligned} \quad (32)$$

which is convex jointly in variables $\tau \geq 0$ and $k \in \mathbb{R}^{|\mathcal{X}|A}$ due to lemma 2 and the fact that $\tau \tilde{K}_\mu^t(s_l, a_l, 1/\tau)$ is the perspective of a convex function. This problem is an *exponential cone program* (see appendix for details) which can be solved efficiently using modern methods [61, 62, 27, 92, 65].

Since problem (32) is convex in the variable k , we could approximate k as $k \approx \Phi x$ for some fixed set of basis vectors $\Phi \in \mathbb{R}^{|\mathcal{X}|A \times m}$ and variable $x \in \mathbb{R}^m$, and then use that in our policy. This is a dimensionality reduction and generalization strategy that, in the no uncertainty case, has theoretical guarantees [25, 82], and has been used successfully in continuous control [109, 110, 66, 67, 60]. This could allow the use of function approximators, such as deep neural networks [55, 54], to be used, where we solve the problem in (32) to find the weights of the last layer.

Algorithm 3 Optimal K-learning for episodic MDPs

Input: MDP $\mathcal{M} = \{\mathcal{X}, \mathcal{A}, R, P, L, \rho\}$,
 initialize $n(s, a) \leftarrow 0$ for all $s, a \in \mathcal{X} \times \mathcal{A}$, $\mathcal{F}_1 = \emptyset$
for episode $t = 1, 2, \dots$ **do**
 solve (32) for optimal β_t and k values
 for step $l = 1, \dots, L$ **do**
 at state s_l sample action a_l with probability $\pi(a_l|s_l) \propto \exp(\beta_t k(s_l, a_l))$
 end for
 update $n(s, a) \leftarrow n(s, a) + 1$ for visited s, a
 update $\mathcal{F}_{t+1} \leftarrow \mathcal{F}_t \cup \{s_l, a_l, r_l, s_{l+1} : l = 1, \dots, L\}$
end for

We can consider the dual problem to (32):

$$\begin{aligned}
 &\text{maximize} && \sum_{l=1}^L \sum_{s_l, a_l} \lambda(s_l, a_l) \mathbf{E}(\mu(s_l, a_l) | \mathcal{F}_t) + \min_{\tau \geq 0} \hat{R}_t(\lambda, \tau) \\
 &\text{subject to} && \sum_{a_l} \lambda(s_l, a_l) = \sum_{s_{l-1}, a_{l-1}} \mathbf{E}(P(s_l | s_{l-1}, a_{l-1}) | \mathcal{F}_t) \lambda(s_{l-1}, a_{l-1}), \quad s_l \in \mathcal{S}_l, \quad l = 2, \dots, L \\
 &&& \sum_{a_1} \lambda(s_1, a_1) = \rho(s_1), \quad s_1 \in \mathcal{S}_1 \\
 &&& \lambda \geq 0
 \end{aligned} \tag{33}$$

which is a concave optimization problem in variable $\lambda \in \mathbb{R}^{|\mathcal{X}|^A}$, where

$$\hat{R}_t(\lambda, \tau) = \sum_{l=1}^L \sum_{s_l, a_l} \lambda(s_l, a_l) \left(\tau \tilde{K}_{\mu - \mathbf{E}(\mu | \mathcal{F}_t)}(s_l, a_l, 1/\tau) + \tau H(\lambda(s_l, \cdot) / \sum_a \lambda(s_l, a)) \right).$$

The solution to the problem, λ^* , can be interpreted as an occupancy measure, where $\lambda^*(s, a)$ is the probability that the agent is in state s and takes action a . Comparing $\hat{R}_t(\lambda^*, \tau)$ to eq. (24) we can see that it upper bounds the expected regret at episode t for any choice of $\tau \geq 0$. Consequently, we can interpret the dual problem as maximizing the expected return plus a term that bounds the expected regret, subject to the (expected) dynamics of the environment. The policy can be recovered as

$$\pi(s, a) = \lambda^*(s, a) / \sum_b \lambda^*(s, b) = \exp(k^*(s, a) / \tau^*) / \sum_b \exp(k^*(s, b) / \tau^*), \tag{34}$$

where (τ^*, k^*) is the solution to the primal problem.

Lemma 12. Assuming strong duality between (32) and (33), the policy given by

$$\pi(s, a) \propto \exp(k^*(s, a) / \tau^*) \propto \lambda^*(s, a),$$

where (τ^*, k^*) is the solution to (32) and where λ^* is the solution to (33), achieves a regret bound at least as good as in theorem 2.

Proof. The argument is essentially the same as in § 2.6. Strong duality implies that the strong max-min property holds for the Lagrangian of this primal-dual problem pair [17, §5.4.1]. If we denote the Lagrangian at τ, k, λ as $\mathcal{L}(\tau, k, \lambda)$, then we have that

$$\mathcal{L}(\tau^*, k^*, \lambda^*) = \min_{\tau, k} \mathcal{L}(\tau, k, \lambda^*) = \min_{\tau} \mathcal{L}(\tau, k^*, \lambda^*)$$

which implies that $\tau^* = \operatorname{argmin}_{\tau \geq 0} \hat{R}_t(\lambda^*, \tau)$, and as such it is the choice of τ that minimizes the expected regret. This implies that the expected regret of policy (34) is upper bounded by that of theorem 2. \square

Note that as the uncertainty tends to zero then the optimal choice of τ tends to zero and we recover the classic primal-dual linear programs that determine the optimal Q-values and policies respectively, see, *e.g.*, [16].

5.4 Numerical example

We compare K-learning against a number of alternative methods in this section. We consider a small tabular MDP called *DeepSea*, adapted from [75], as shown in figure 4. This MDP can be visualized as an $L \times L$ grid, the agent starts at the top row and leftmost column. At each time-period the agent chooses from actions ‘left’ or ‘right’, and then is advanced down one row and moved one column to the left or to the right. If the agent chose action ‘left’ then it is moved one column to the left, and if it chose ‘right’ then it moves one column to the right with probability $1 - p$ and to the left with probability p . If the agent is against the edge of the grid and the result of the action would be to move ‘into’ the edge then it remains in that column (but still moves down one row).

At any state, after choosing action ‘left’ the agent receives a reward sampled from $\mathcal{N}(0, 1)$ and after choosing right the agent receives a reward sampled from $\mathcal{N}(-\epsilon, 1)$ for some $\epsilon > 0$. At the bottom rightmost corner of the grid the agent receives a reward sampled from $\mathcal{N}(1, 1)$. All parameters are chosen so that the optimal policy is to try to get to the bottom right state, if it is still possible, and move left otherwise. In this setup the agent must incur negative expected reward for several time-steps before receiving the positive reward.

Though this is a very simple toy example, the results are instructive. We present the graph of results for two different MDP sizes, $L = 10$ and $L = 30$, in figure 5 and in table 1. Each algorithm was run from five different seeds for each MDP and the numbers reported are an average. For a grid-size of $L = 30$ a random agent that takes each action with probability 0.5 would have a probability of $2^{-30} \approx 10^{-9}$ of reaching the bottom right (without considering the fact that taking action right does not always result in moving right). Therefore efficient exploration in this example requires deep and directed exploration [64] since local jittering, as in ϵ -greedy, will take a very long time to learn the optimal policy. The Bayesian algorithms require a prior, we chose the prior over the mean rewards at each state-action to be $\mathcal{N}(0, 1)$, and we chose the prior over transition probabilities to be Dirichlet with parameter one for each state at the next time-step, *i.e.*, uniform.

For the $L = 10$ case PSRL [74] (posterior sampling for RL) is the best performing algorithm, though for $L = 30$ K-learning with the optimal choice of β performs best. K-learning with the β_t schedule and greedy K-learning (described in the appendix) are the next best performers. One thing to note is that the rate at which the regret is increasing for these non-optimal variants of K-learning is higher than for PSRL and so it would seem that eventually PSRL will have lower regret than these techniques, however the same does not appear to be true for K-learning with the optimal choice of β which looks to have a similar asymptotic increase in regret as PSRL, despite only using a variational approximation to the posterior. Although K-learning performs best in this case, we make no claims that K-learning is better than PSRL overall (indeed they have essentially the same regret bound). One possible explanation for the better performance of K-learning in this case is that K-learning is explicitly risk-seeking, and so will prioritize actions that have higher uncertainty. In this particular example it means that K-learning will seek out the bottom right corner which has been constructed to have a high reward on average.

Other techniques perform much worse on these examples. RLSVI [75] is an approximation to PSRL that achieves a good theoretical regret bound, however in this case it does not perform well, though it has several tuning parameters that could potentially be tweaked further. Optimism-in-the-face-of-uncertainty techniques UCBVI [10] and UCRL2 [36] perform poorly because they take a very long time to decrease their confidence sets. Their performance could also possibly be improved by tuning their scaling parameters. For ϵ -greedy linear regret is expected, because the choice of ϵ is fixed at 0.05.

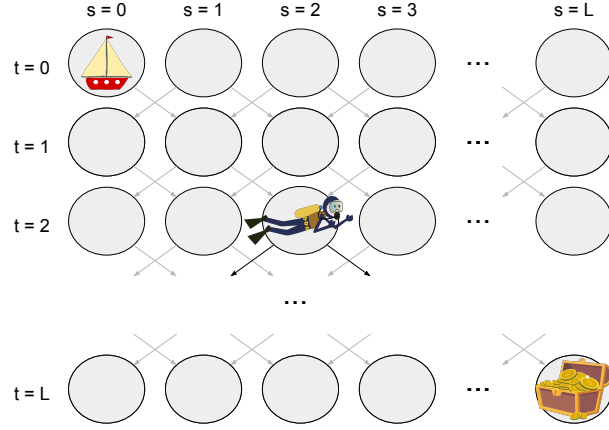


Figure 4: The DeepSea MDP.

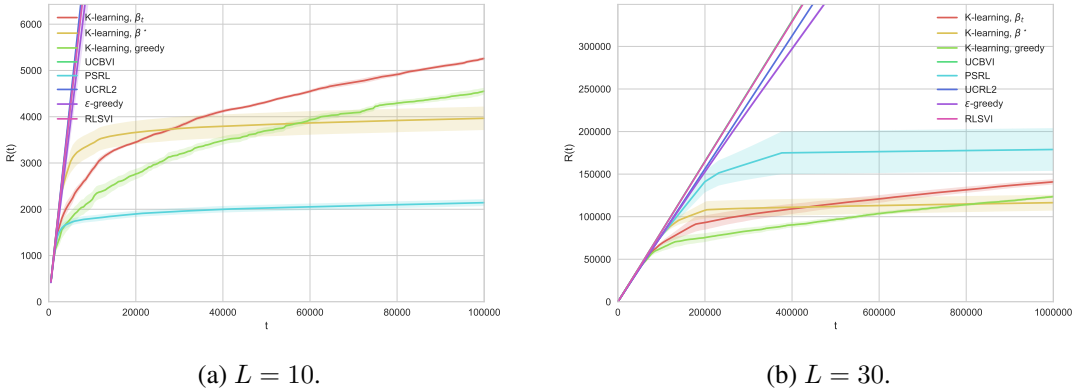


Figure 5: Regret over time for tested algorithms on DeepSea MDP.

5.5 Connections with other methods

Other RL methods. In the case of no uncertainty it can be shown that K -learning is essentially equivalent to standard max-Q learning. In this case the optimal $\beta_t = \infty$ for all t and the bonus afforded to each action is zero, and so the K -learning update eq. (31) simply reduces to the max-Q-learning update eq. (20).

$L = 10$		$L = 30$	
K-learning, β_t	2.45	K-learning, β_t	1.21
K-learning, β^*	1.85	K-learning, β^*	1.00
K-learning, greedy	2.12	K-learning, greedy	1.06
UCBVI	37.01	UCBVI	7.08
PSRL	1.00	PSRL	1.54
UCRL2	40.86	UCRL2	6.70
ϵ -greedy	31.82	ϵ -greedy	6.12
RLSVI	19.24	RLSVI	7.05

Table 1: Regret for Deep Sea examples, as a multiple of lowest regret.

If we remove the bonus term from eq. (31) we recover soft Q-learning, which is Q-learning where the max operator has been replaced by a ‘soft’ max. Although soft Q-learning has no performance guarantees it has been shown to perform well in practice. Due to lemma 3 we can rewrite the log-sum-exp term as the expected value under policy π where the rewards are given an entropy regularizer, and so soft Q-learning can also be interpreted as entropy regularized SARSA.

In [63] the authors showed that any entropy-regularized actor-critic algorithm can be interpreted as a value function learning algorithm, where the log-policy is estimating the advantage function. The authors of [89] showed that soft Q-learning is equivalent (in expectation) to a particular form of entropy regularized actor-critic, which suggests a connection between K-learning and actor-critic algorithms, at least if implemented online. Proving a regret bound in the online (*i.e.*, model-free) case for K-learning is an open problem.

Count based exploration. Count based exploration affords a bonus to the reward at each state-action based on the number of times it has been taken before. In MBIE-EB the authors take the bonus to be proportional to $1/\sqrt{n(s, a)}$ and in BEB the authors take it to be proportional to $1/n(s, a)$. In contrast our approach gives a bonus proportional to $\sqrt{t}/n(s, a)$ where t is the episode counter. Note that since $t \geq n(s, a)$ due to assumption 2 it implies that the bonus is at least $1/\sqrt{n(s, a)}$. This is necessary to achieve a Bayesian expected regret bound of $O(\sqrt{T})$, which these other works do not achieve, instead suffering from linear expected regret. Indeed in [24] it was shown that in order to achieve a sublinear expected regret bound every state must be visited infinitely often, which precludes a PAC guarantee. In the non-tabular case methods that estimate the pseudo-count have proven successful in some instances, indicating a possibility of combining K-learning with pseudo-counts in a deep RL setting [94, 11, 97, 80, 13, 79].

Mirror descent. Mirror descent is an iterative algorithm to minimize a convex function

$$x_{t+1} = \operatorname{argmin}_{x \in \mathcal{D}} (\eta x^T g_t + D_\Phi(x, x_t))$$

where g_t is a (possibly stochastic) subgradient of the convex function to be minimized, D_Φ is a Bregman divergence, \mathcal{D} is a convex set and $\eta > 0$ is a stepsize [59, 12]. One particularly commonly used divergence is the KL-divergence, *i.e.*, $D_\Phi(x, y) = D_{\text{KL}}(x, y) = \sum x_i \log(x_i/y_i)$. We show here that the policy update in lemma 3 is performing a variant of mirror descent. For ease of exposition we will take $\beta_t = \beta$ a constant fixed for all episodes (we get essentially the same regret bound for a fixed $\beta = \beta_T$ as for a changing β , but

we lose the anytime nature of the algorithm). Consider the update to the K-function that occurs after time t , i.e., let $\delta k_t(s, a) = k_{t+1}(s, a) - k_t(s, a)$. According to lemma 3 the policy at any state at each episode satisfies (suppressing the dependence on state)

$$\begin{aligned}\pi_{t+1} &= \operatorname{argmax}_{\pi \in \Pi} (\beta \pi^T (k_t + \delta k_t) + H(\pi)) \\ &= \operatorname{argmax}_{\pi \in \Pi} (\beta \pi^T \delta k_t + \pi^T (\beta k_t - \log \pi)) \\ &= \operatorname{argmax}_{\pi \in \Pi} (\beta \pi^T \delta k_t + \pi^T (\log \pi_t - \log \pi)) \\ &= \operatorname{argmin}_{\pi \in \Pi} (-\beta \pi^T \delta k_t + D_{\text{KL}}(\pi, \pi_t)),\end{aligned}$$

where we used the fact that $\pi_t \propto \exp \beta k^t$. In the online convex learning case, where the learner receives a sequence of functions and at time t uses the subgradient of the function received at that time, mirror descent has a bound on regret of $\tilde{O}(\sqrt{T})$, which matches our bound [93]. The relationship here between mirror descent with a KL-penalty and K-learning also suggests a relationship to TRPO, PPO, and natural policy gradient [90, 91, 38], which update the policy along a gradient subject to a KL-divergence penalty or constraint.

6 Conclusions

In this paper we derived a new Bellman update rule and associated exploration policy. We refer to this algorithm as K-learning, and it guarantees a Bayesian expected regret bound for episodic MDPs of $\tilde{O}(L^{3/2} \sqrt{SAT})$ where L is the time horizon, S is the number of states, A is the number of actions, and T is the total number of elapsed time-steps. This bound is only a factor of L larger than the established lower bound. In the presented numerical examples K-learning performed well when compared to other techniques from the literature.

7 Acknowledgments

I would like to thank Ian Osband, Remi Munos, Vlad Mnih, Pedro Ortega, and Yee Whye Teh for their support, valuable discussions, and clear insights.

References

- [1] A. E. ABBAS, *Invariant utility functions and certain equivalent transformations*, Decision Analysis, 4 (2007), pp. 17–31.
- [2] A. ABDOLMALEKI, J. T. SPRINGENBERG, Y. TASSA, R. M. N. HEES, AND M. RIEDMILLER, *Maximum a posteriori policy optimisation*, in International Conference on Learning Representations (ICLR), 2018.
- [3] S. AGRAWAL AND N. GOYAL, *Near-optimal regret bounds for thompson sampling*, Journal of the ACM (JACM), 64 (2017), p. 30.
- [4] M. ARAYA, O. BUFFET, AND V. THOMAS, *Near-optimal BRL using optimistic local transitions*, in Proceedings of the 29th International Conference on Machine Learning (ICML), 2012.

- [5] J.-Y. AUDIBERT AND S. BUBECK, *Minimax policies for adversarial and stochastic bandits*, in COLT, 2009, pp. 217–226.
- [6] P. AUER, N. CESA-BIANCHI, AND P. FISCHER, *Finite-time analysis of the multiarmed bandit problem*, Machine learning, 47 (2002), pp. 235–256.
- [7] P. AUER, N. CESA-BIANCHI, Y. FREUND, AND R. E. SCHAPIRE, *The nonstochastic multiarmed bandit problem*, SIAM journal on computing, 32 (2002), pp. 48–77.
- [8] P. AUER AND R. ORTNER, *Logarithmic online regret bounds for undiscounted reinforcement learning*, in NIPS, vol. 19, 2006, pp. 49–56.
- [9] M. G. AZAR, V. GÓMEZ, AND H. J. KAPPEN, *Dynamic policy programming*, Journal of Machine Learning Research, 13 (2012), pp. 3207–3245.
- [10] M. G. AZAR, I. OSBAND, AND R. MUNOS, *Minimax regret bounds for reinforcement learning*, in International Conference on Machine Learning, 2017, pp. 263–272.
- [11] A. G. BARTO, *Intrinsic motivation and reinforcement learning*, in Intrinsically motivated learning in natural and artificial systems, Springer, 2013, pp. 17–47.
- [12] A. BECK AND M. TEBoulLE, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, 31 (2003), pp. 167–175.
- [13] M. BELLEMARE, S. SRINIVASAN, G. OSTROVSKI, T. SCHAUL, D. SAXTON, AND R. MUNOS, *Unifying count-based exploration and intrinsic motivation*, in Advances in Neural Information Processing Systems, 2016, pp. 1471–1479.
- [14] M. G. BELLEMARE, W. DABNEY, AND R. MUNOS, *A distributional perspective on reinforcement learning*, in International Conference on Machine Learning, 2017, pp. 449–458.
- [15] R. BELLMAN, *Dynamic programming*, Princeton University Press, 1957.
- [16] D. P. BERTSEKAS, *Dynamic programming and optimal control*, vol. 1, Athena Scientific, 2005.
- [17] S. BOYD AND L. VANDENBERGHE, *Convex optimization*, Cambridge university press, 2004.
- [18] R. I. BRAFMAN AND M. TENNENHOLTZ, *R-max: A general polynomial time algorithm for near-optimal reinforcement learning*, Journal of Machine Learning Research, 3 (2002), pp. 213–231.
- [19] S. BUBECK AND N. CESA-BIANCHI, *Regret analysis of stochastic and nonstochastic multi-armed bandit problems*, Foundations and Trends® in Machine Learning, 5 (2012), pp. 1–122.
- [20] N. CESA-BIANCHI AND G. LUGOSI, *Prediction, learning, and games*, Cambridge university press, 2006.
- [21] N. CHENTANEZ, A. G. BARTO, AND S. P. SINGH, *Intrinsically motivated reinforcement learning*, in Advances in neural information processing systems, 2005, pp. 1281–1288.
- [22] T. M. COVER AND J. A. THOMAS, *Elements of information theory*, John Wiley & Sons, 2012.

- [23] C. DANN AND E. BRUNSKILL, *Sample complexity of episodic fixed-horizon reinforcement learning*, in Advances in Neural Information Processing Systems, 2015, pp. 2818–2826.
- [24] C. DANN, T. LATTIMORE, AND E. BRUNSKILL, *Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning*, in Advances in Neural Information Processing Systems, 2017, pp. 5717–5727.
- [25] D. P. DE FARIAS AND B. VAN ROY, *The linear programming approach to approximate dynamic programming*, Operations research, 51 (2003), pp. 850–865.
- [26] S. DE ROOIJ, T. VAN ERVEN, P. D. GRÜN WALD, AND W. M. KOOLEN, *Follow the leader if you can, hedge if you must*, The Journal of Machine Learning Research, 15 (2014), pp. 1281–1316.
- [27] A. DOMAHIDI, E. CHU, AND S. BOYD, *ECOS: An SOCP solver for embedded systems*, in European Control Conference (ECC), 2013, pp. 3071–3076.
- [28] R. FOX, A. PAKMAN, AND N. TISHBY, *Taming the noise in reinforcement learning via soft updates*. arXiv preprint arXiv:1207.4708, 2015.
- [29] M. GHAVAMZADEH, S. MANNOR, J. PINEAU, AND A. TAMAR, *Bayesian reinforcement learning: A survey*, Foundations and Trends® in Machine Learning, 8 (2015), pp. 359–483.
- [30] J. C. GITTINS, *Bandit processes and dynamic allocation indices*, Journal of the Royal Statistical Society. Series B (Methodological), (1979), pp. 148–177.
- [31] T. HAARNOJA, H. TANG, P. ABBEEL, AND S. LEVINE, *Reinforcement learning with deep energy-based policies*, in Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.
- [32] J. HADAR AND W. R. RUSSELL, *Rules for ordering uncertain prospects*, The American economic review, 59 (1969), pp. 25–34.
- [33] R. HOUTHOOFT, X. CHEN, Y. DUAN, J. SCHULMAN, F. DE TURCK, AND P. ABBEEL, *VIME: Variational information maximizing exploration*, in Advances in Neural Information Processing Systems, 2016, pp. 1109–1117.
- [34] R. A. HOWARD, *Value of information lotteries*, IEEE Transactions on Systems Science and Cybernetics, 3 (1967), pp. 54–60.
- [35] R. A. HOWARD AND J. E. MATHESON, *Risk-sensitive markov decision processes*, Management science, 18 (1972), pp. 356–369.
- [36] T. JAKSCH, R. ORTNER, AND P. AUER, *Near-optimal regret bounds for reinforcement learning*, Journal of Machine Learning Research, 11 (2010), pp. 1563–1600.
- [37] B. JORGENSEN, *The theory of dispersion models*, CRC Press, 1997.
- [38] S. KAKADE, *A natural policy gradient*, in Advances in Neural Information Processing Systems, vol. 14, 2001, pp. 1531–1538.
- [39] S. M. KAKADE, *On the sample complexity of reinforcement learning*, PhD thesis, University of London London, England, 2003.

- [40] E. KAUFMANN, O. CAPPÉ, AND A. GARIVIER, *On Bayesian upper confidence bounds for bandit problems*, in Artificial Intelligence and Statistics, 2012, pp. 592–600.
- [41] M. KEARNS AND S. SINGH, *Near-optimal reinforcement learning in polynomial time*, Machine Learning, 49 (2002), pp. 209–232.
- [42] M. G. KENDALL, *The advanced theory of statistics.*, Charles Griffin and Co., Ltd., London, 1946.
- [43] A. S. KLYUBIN, D. POLANI, AND C. L. NEHANIV, *Empowerment: A universal agent-centric measure of control*, in Evolutionary Computation, 2005. The 2005 IEEE Congress on, vol. 1, IEEE, 2005, pp. 128–135.
- [44] J. Z. KOLTER AND A. Y. NG, *Near-Bayesian exploration in polynomial time*, in Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 513–520.
- [45] T. LATTIMORE AND M. HUTTER, *Pac bounds for discounted mdps*, in International Conference on Algorithmic Learning Theory, Springer, 2012, pp. 320–334.
- [46] S. LEVINE, *Reinforcement learning and control as probabilistic inference: Tutorial and review*, arXiv preprint arXiv:1805.00909, (2018).
- [47] Z. C. LIPTON, J. GAO, L. LI, X. LI, F. AHMED, AND L. DENG, *Efficient exploration for dialogue policy learning with BBQ networks & replay buffer spiking*, arXiv preprint arXiv:1608.05081, (2016).
- [48] Y. LIU, R. GOODWIN, AND S. KOENIG, *Risk-averse auction agents*, in Proceedings of the second international joint conference on Autonomous agents and multiagent systems, ACM, 2003, pp. 353–360.
- [49] M. LOPES, T. LANG, M. TOUSSAINT, AND P.-Y. OUDEYER, *Exploration in model-based reinforcement learning by empirically estimating learning progress*, in Advances in Neural Information Processing Systems, 2012, pp. 206–214.
- [50] C. J. MADDISON, D. LAWSON, G. TUCKER, N. HEES, A. DOUCET, A. MNIH, AND Y. W. TEH, *Particle value functions*, arXiv preprint arXiv:1703.05820, (2017).
- [51] O. MARCHAL AND J. ARBEL, *On the sub-Gaussianity of the Beta and Dirichlet distributions*. arXiv preprint arXiv:1705.00048, 2017.
- [52] S. I. MARCUS, E. FERNÁNDEZ-GAUCHERAND, D. HERNÁNDEZ-HERNANDEZ, S. CORALUPPI, AND P. FARD, *Risk sensitive markov decision processes*, in Systems and control in the twenty-first century, Springer, 1997, pp. 263–279.
- [53] O. MIHATSCH AND R. NEUNEIER, *Risk-sensitive reinforcement learning*, Machine learning, 49 (2002), pp. 267–290.
- [54] V. MNIH, A. P. BADIA, M. MIRZA, A. GRAVES, T. LILICRAP, T. HARLEY, D. SILVER, AND K. KAVUKCUOGLU, *Asynchronous methods for deep reinforcement learning*, in Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016, pp. 1928–1937.

- [55] V. MNIH, K. KAVUKCUOGLU, D. SILVER, A. A. RUSU, J. VENESS, M. G. BELLEMARE, A. GRAVES, M. RIEDMILLER, A. K. FIDJELAND, G. OSTROVSKI, S. PETERSEN, C. BEATTIE, A. SADIK, I. ANTONOGLU, H. KING, D. KUMARAN, D. WIERSTRA, S. LEGG, AND D. HASSABIS, *Human-level control through deep reinforcement learning*, *Nature*, 518 (2015), pp. 529–533.
- [56] S. MOHAMED AND D. J. REZENDE, *Variational information maximisation for intrinsically motivated reinforcement learning*, in *Advances in neural information processing systems*, 2015, pp. 2125–2133.
- [57] R. MUNOS, *From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning*, *Foundations and Trends® in Machine Learning*, 7 (2014), pp. 1–129.
- [58] O. NACHUM, M. NOROUZI, K. XU, AND D. SCHUURMANS, *Bridging the gap between value and policy based reinforcement learning*, in *Advances in Neural Information Processing Systems*, 2017, pp. 2772–2782.
- [59] A. NEMIROVSKI AND D. B. YUDIN, *Problem complexity and method efficiency in optimization*, Wiley, 1983.
- [60] B. O’DONOGHUE, *Suboptimal control policies via convex optimization*, PhD thesis, Stanford University, 2012.
- [61] B. O’DONOGHUE, E. CHU, N. PARIKH, AND S. BOYD, *Conic optimization via operator splitting and homogeneous self-dual embedding*, *Journal of Optimization Theory and Applications*, 169 (2016), pp. 1042–1068.
- [62] ———, *SCS: Splitting conic solver, version 2.0.2*. <https://github.com/cvxgrp/scs>, Nov. 2017.
- [63] B. O’DONOGHUE, R. MUNOS, K. KAVUKCUOGLU, AND V. MNIH, *Combining policy gradient and Q-learning*, in *International Conference on Learning Representations (ICLR)*, 2017.
- [64] B. O’DONOGHUE, I. OSBAND, R. MUNOS, AND V. MNIH, *The uncertainty Bellman equation and exploration*, arXiv preprint arXiv:1709.05380, (2017).
- [65] B. O’DONOGHUE, G. STATHOPOULOS, AND S. BOYD, *A splitting method for optimal control*, *IEEE Transactions on Control Systems Technology*, 21 (2013), pp. 2432–2442.
- [66] B. O’DONOGHUE, Y. WANG, AND S. BOYD, *Min-max approximate dynamic programming*, in *Computer-Aided Control System Design (CACSD)*, 2011 IEEE International Symposium on, IEEE, 2011, pp. 424–431.
- [67] ———, *Iterated approximate value functions*, in *Control Conference (ECC)*, 2013 European, IEEE, 2013, pp. 3882–3888.
- [68] P. A. ORTEGA AND D. A. BRAUN, *Information, utility and bounded rationality*, in *International Conference on Artificial General Intelligence*, Springer, 2011, pp. 269–274.
- [69] ———, *Free energy and the generalized optimality equations for sequential decision making*, in *European Workshop on Reinforcement Learning*, 2012.

- [70] P. A. ORTEGA AND D. A. BRAUN, *Thermodynamics as a theory of decision-making with information-processing costs*, Proc. R. Soc. A, 469 (2013), p. 20120683.
- [71] P. A. ORTEGA, D. A. BRAUN, J. DYER, K.-E. KIM, AND N. TISHBY, *Information-theoretic bounded rationality*, arXiv preprint arXiv:1512.06789, (2015).
- [72] P. A. ORTEGA JR, *A unified framework for resource-bounded autonomous agents interacting with unknown environments*, PhD thesis, University of Cambridge, 2011.
- [73] I. OSBAND, C. BLUNDELL, A. PRITZEL, AND B. VAN ROY, *Deep exploration via bootstrapped DQN*, in Advances In Neural Information Processing Systems, 2016, pp. 4026–4034.
- [74] I. OSBAND, D. RUSSO, AND B. VAN ROY, *(More) efficient reinforcement learning via posterior sampling*, in Advances in Neural Information Processing Systems, 2013, pp. 3003–3011.
- [75] I. OSBAND, D. RUSSO, Z. WEN, AND B. VAN ROY, *Deep exploration via randomized value functions*, arXiv preprint arXiv:1703.07608, (2017).
- [76] I. OSBAND AND B. VAN ROY, *Gaussian-Dirichlet posterior dominance in sequential learning*, arXiv preprint arXiv:1702.04126, (2017).
- [77] ———, *Why is posterior sampling better than optimism for reinforcement learning*, in Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.
- [78] I. OSBAND, B. VAN ROY, AND Z. WEN, *Generalization and exploration via randomized value functions*, arXiv preprint arXiv:1402.0635, (2014).
- [79] G. OSTROVSKI, M. G. BELLEMARE, A. V. D. OORD, AND R. MUNOS, *Count-based exploration with neural density models*, arXiv preprint arXiv:1703.01310, (2017).
- [80] P.-Y. OUDEYER, F. KAPLAN, AND V. V. HAFNER, *Intrinsic motivation systems for autonomous mental development*, IEEE transactions on evolutionary computation, 11 (2007), pp. 265–286.
- [81] J. PFANZAG, *A general theory of measurement applications to utility*, Naval Research Logistics (NRL), 6 (1959), pp. 283–294.
- [82] W. B. POWELL, *Approximate Dynamic Programming: Solving the curses of dimensionality*, vol. 703, John Wiley & Sons, 2007.
- [83] J. W. PRATT, *Risk aversion in the small and in the large*, in Stochastic Optimization Models in Finance, Elsevier, 1975, pp. 115–130.
- [84] M. L. PUTERMAN, *Markov decision processes: Discrete stochastic dynamic programming*, John Wiley & Sons, 2014.
- [85] H. RAIFFA, *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, Addison Wesley, 1968.
- [86] P. H. RICHEMOND AND B. MAGINNIS, *A short variational proof of equivalence between policy gradients and soft Q learning*, arXiv preprint arXiv:1712.08650, (2017).

- [87] D. RUSSO AND B. VAN ROY, *Learning to optimize via posterior sampling*, Mathematics of Operations Research, 39 (2014), pp. 1221–1243.
- [88] ———, *An information-theoretic analysis of Thompson sampling*, The Journal of Machine Learning Research, 17 (2016), pp. 2442–2471.
- [89] J. SCHULMAN, P. ABBEEL, AND X. CHEN, *Equivalence between policy gradients and soft Q-learning*, arXiv preprint arXiv:1704.06440, (2017).
- [90] J. SCHULMAN, S. LEVINE, P. ABBEEL, M. JORDAN, AND P. MORITZ, *Trust region policy optimization*, in Proceedings of The 32nd International Conference on Machine Learning, 2015, pp. 1889–1897.
- [91] J. SCHULMAN, F. WOLSKI, P. DHARIWAL, A. RADFORD, AND O. KLIMOV, *Proximal policy optimization algorithms*, arXiv preprint arXiv:1707.06347, (2017).
- [92] S. A. SERRANO, *Algorithms for unsymmetric cone optimization and an implementation for problems with the exponential cone*, PhD thesis, Stanford University, 2015.
- [93] S. SHALEV-SHWARTZ, *Online learning and online convex optimization*, Foundations and Trends® in Machine Learning, 4 (2012), pp. 107–194.
- [94] S. P. SINGH, A. G. BARTO, AND N. CHENTANEZ, *Intrinsically motivated reinforcement learning*, in NIPS, vol. 17, 2004, pp. 1281–1288.
- [95] M. SION, *On general minimax theorems*, Pacific Journal of mathematics, 8 (1958), pp. 171–176.
- [96] M. J. SOBEL, *The variance of discounted Markov decision processes*, Journal of Applied Probability, 19 (1982), pp. 794–802.
- [97] B. C. STADIE, S. LEVINE, AND P. ABBEEL, *Incentivizing exploration in reinforcement learning with deep predictive models*, arXiv preprint arXiv:1507.00814, (2015).
- [98] A. STREHL AND M. LITTMAN, *Exploration via model-based interval estimation*, 2004.
- [99] A. L. STREHL, L. LI, E. WIEWIORA, J. LANGFORD, AND M. L. LITTMAN, *PAC model-free reinforcement learning*, in Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 881–888.
- [100] A. L. STREHL AND M. L. LITTMAN, *A theoretical analysis of model-based interval estimation*, in Proceedings of the 22nd international conference on Machine learning, ACM, 2005, pp. 856–863.
- [101] M. STRENS, *A Bayesian framework for reinforcement learning*, in ICML, 2000, pp. 943–950.
- [102] R. SUTTON AND A. BARTO, *Reinforcement Learning: an Introduction*, MIT Press, 1998.
- [103] I. SZITA AND C. SZEPESVÁRI, *Model-based reinforcement learning with nearly tight exploration complexity bounds*, in Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 1031–1038.
- [104] A. TAMAR, D. DI CASTRO, AND S. MANNOR, *Learning the variance of the reward-to-go*, Journal of Machine Learning Research, 17 (2016), pp. 1–36.

- [105] W. R. THOMPSON, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika, 25 (1933), pp. 285–294.
- [106] J. VON NEUMANN, *Zur theorie der gesellschaftsspiele*, Mathematische annalen, 100 (1928), pp. 295–320.
- [107] J. VON NEUMANN AND O. MORGENSTERN, *Theory of games and economic behavior (commemorative edition)*, Princeton university press, 2007.
- [108] M. J. WAINWRIGHT AND M. I. JORDAN, *Graphical models, exponential families, and variational inference*, Foundations and Trends® in Machine Learning, 1 (2008), pp. 1–305.
- [109] Y. WANG AND S. BOYD, *Performance bounds for linear stochastic control*, Systems & Control Letters, 58 (2009), pp. 178–182.
- [110] Y. WANG, B. O’DONOGHUE, AND S. BOYD, *Approximate dynamic programming via iterated Bellman inequalities*, International Journal of Robust and Nonlinear Control, 25 (2015), pp. 1472–1496.
- [111] M. WHITE AND A. WHITE, *Interval estimation for reinforcement-learning algorithms in continuous-state domains*, in Advances in Neural Information Processing Systems, 2010, pp. 2433–2441.
- [112] R. J. WILLIAMS AND J. PENG, *Function optimization using connectionist reinforcement learning algorithms*, Connection Science, 3 (1991), pp. 241–268.
- [113] B. D. ZIEBART, *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*, Carnegie Mellon University, 2010.

8 Appendix

8.1 Greedy K-learning for Bandits

Algorithm 4 Greedy K-learning for multi-armed bandits

Input: multi-armed bandit $\{A, R\}$, uncertainty parameter σ
 initialize $n_a \leftarrow 0, a = 1, \dots, A, \mathcal{F}_1 \leftarrow \emptyset$
for round $t = 1, 2, \dots$, **do**
 calculate $\beta_t = \sqrt{\frac{4t \log A}{\sigma^2 A(1+\log t)}}$
 calculate posterior mean of rewards $\hat{\mu} = \mathbf{E}(\mu|\mathcal{F}_t)$
 calculate k values for $a = 1, \dots, A$: $k_a = \hat{\mu}_a + \frac{\sigma^2 \beta_t}{2(1+n_a)}$
 action $a_t = \operatorname{argmax} k_a$
 update $n_{a_t} \leftarrow n_{a_t} + 1, \mathcal{F}_{t+1} \leftarrow \mathcal{F}_t \cup \{a_t, r_t\}$
end for

8.2 Greedy K-learning for MDPs

Algorithm 5 Greedy K-learning for episodic MDPs

Input: MDP $\mathcal{M} = \{\mathcal{X}, \mathcal{A}, R, P, L, \rho\}$,
 initialize $n(s, a) \leftarrow 0$ for all $s, a \in \mathcal{X} \times \mathcal{A}, \mathcal{F}_1 \leftarrow \emptyset$
for episode $t = 1, 2, \dots$ **do**
 calculate $\beta_t = \sqrt{\frac{4tL \log A}{(\sigma_r^2 + L^2)A|\mathcal{X}|(1+\log t)}}$
 calculate $\hat{\mu} = \mathbf{E}(\mu|\mathcal{F}_t)$ and transition operator $\hat{\mathbf{E}}_{s_{l+1}}$ via $\mathbf{E}(P|\mathcal{F}_t)$
 set $k(s_{L+1}, \cdot) = 0$
 for step $l = L, \dots, 1$ **do**

$$k(s_l, a_l) = \hat{\mu}(s_l, a_l) + \frac{(\sigma_r^2 + L^2)\beta_t}{2(1 + n_t(s_l, a_l))} + \hat{\mathbf{E}}_{s_{l+1}} \max_{a_{l+1}} k(s_{l+1}, a_{l+1}) \quad (35)$$

for each $(s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}$
 end for
 for step $l = 1, \dots, L$ **do**
 at state s_l calculate action $a_l = \operatorname{argmax} k(s_l, \cdot)$
 end for
 update $n(s, a) \leftarrow n(s, a) + 1$ for visited s, a
 update $\mathcal{F}_{t+1} \leftarrow \mathcal{F}_t \cup \{s_l, a_l, r_l, s_{l+1} : l = 1, \dots, L\}$
end for

8.3 The perspective preserves strict convexity

Lemma 13. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a strictly convex function, then $g : \mathbb{R}_+ \rightarrow \mathbb{R}$ defined by $g(\tau) = \tau f(1/\tau)$ is also strictly convex.

Proof. This follows from a straightforward application of Jensen's inequality, let $\theta \in (0, 1)$,

$$\begin{aligned}
g(\theta x + (1 - \theta)y) &= (\theta x + (1 - \theta)y) f\left(\frac{1}{\theta x + (1 - \theta)y}\right) \\
&= (\theta x + (1 - \theta)y) f\left(\frac{\theta x}{\theta x + (1 - \theta)y} \frac{1}{x} + \frac{(1 - \theta)y}{\theta x + (1 - \theta)y} \frac{1}{y}\right) \\
&< (\theta x + (1 - \theta)y) \left(\frac{\theta x}{\theta x + (1 - \theta)y} f\left(\frac{1}{x}\right) + \frac{(1 - \theta)y}{\theta x + (1 - \theta)y} f\left(\frac{1}{y}\right) \right) \\
&= \theta g(x) + (1 - \theta)g(y).
\end{aligned}$$

□

8.4 Conic representation of perspective of log-sum-exp

Here we show the exponential cone formulation of the epigraph for $f(x, t) = t \log \sum_{i=1}^m \exp(x_i/t)$ which, as we have already shown, is jointly convex in x and t . The *exponential cone* $\mathcal{C}_{\text{exp}} \subset \mathbb{R}^3$ is defined as

$$\mathcal{C}_{\text{exp}} = \{(x, y, z) : y \exp(x/y) \leq z, y \geq 0\}.$$

We want a conic formulation for the set $\{(x, t, y) : f(x, t) \leq y\}$. We can rewrite the inequality as $\sum_i \exp(x_i/t) \leq \exp(y/t)$, which is equivalent to

$$t \exp((x_i - y)/t) \leq v_i, \quad \sum_{i=1}^m v_i = t$$

from which we get the conic representation

$$(x_i - y, t, v_i) \in \mathcal{C}_{\text{exp}}, \quad i = 1, \dots, m, \quad \sum_{i=1}^m v_i = t.$$