Possibilistic Q-learning: Uncertainty Modelling for Tuning-Free Optimism

Jake Thomas
Warwick Mathematical Institute
University of Warwick
Coventry, UK
j.h.thomas@warwick.ac.uk

Jeremie Houssineau
Division of Mathematical Sciences
Nanyang Technological University
Singapore
jeremie.houssineau@ntu.edu.sg

Abstract

This paper introduces a new framework for model-based reinforcement learning that builds upon a unique integration of possibility theory and probability theory, and utilises this framework to develop novel algorithms. Unlike traditional methods that rely solely on probability theory to manage all forms of uncertainty, our approach differentiates aleatoric and epistemic uncertainties. By leveraging possibility theory, the proposed method represents epistemic uncertainty in the transition and reward dynamics, accurately capturing the two types of uncertainty. We explore the opportunities created by this framework at the methodological level and focus on one particular instance which yields an algorithm that is free of hyperparameters and equipped with an inherent exploration mechanism that naturally balances the exploration-exploitation trade-off. The proposed algorithm, named Possibilistic Q-learning, utilises maximum expected Q-values, which are reasonable upper bounds of the Q-value given the current level of epistemic uncertainty. The theoretical approach is supported by strong experimental results on a variety of tabular problems.

1 Introduction

Reinforcement learning, where an agent iteratively interacts with an environment to select actions that maximise its cumulative reward, is one of the core problems in artificial intelligence. Sample efficiency is a key challenge in reinforcement learning as the success of existing approaches depend on many trials, which may not be available in practical applications. Model-based reinforcement learning (Sutton, 1990; Atkeson, 1997; Berkenkamp et al., 2017; Moerland et al., 2023) has shown significant potential to improve sample efficiency, either directly (Sun et al., 2019; Zhang et al., 2020) or in support of model-free approaches (Feinberg et al., 2018).

Model-based approaches have to be handled carefully as the uncertainty within the model can quickly compound and produce inaccurate values and policies (Amini et al., 2020). Therefore, faithfully modelling the uncertainty within the environment model is vital for creating effective model-based reinforcement learning algorithms. To this end, we consider a particular combination (Houssineau, 2018) of possibility theory (Dubois & Prade, 2015) and probability theory in order to differentiate between epistemic and random uncertainty, with epistemic uncertainty referring to uncertainty about fixed quantities (such as model parameters) caused by a lack of information, and random uncertainty referring to uncertainty caused by truly random, unpredictable objects (such as random transitions and rewards). The importance of the distinction between these two types of uncertainty is well understood in Machine Learning (Amini et al., 2020; Hüllermeier & Waegeman, 2021). The standard statistical approach of modelling both with probability theory can limit the benefits of such a distinction in practice. For example, the properties of the probabilistic notion of expected value often do not match with the specific behaviour of epistemic uncertainty. Our

approach aims to refine the management of uncertainties inherent in the Temporal Difference (TD) update process, utilising "optimism in the face of uncertainty" (Kaelbling et al., 1996) to address the overestimation bias of the TD update and guide exploration.

In this paper, we introduce a framework for model-based reinforcement learning that explicitly models the deterministic transition and reward parameters as well as the random rewards and transitions. In Section 2 we introduce a novel estimation method for the value function of a Markov reward process and apply this experimentally to a 2-dimensional GridWorld example. This is followed by the full framework being introduced in Section 3, this section also includes a new algorithm and exploration strategies that follow naturally from the framework. Finally, Section 4 demonstrates how our approach can be applied to a variety of GridWorld environments and provides experimental performance results before concluding in Section 5.

1.1 Related Work

Temporal Difference value-based methods, such as Q-learning and SARSA (Sutton & Barto, 1998), form the core of traditional reinforcement learning approaches. However, these methods are known to suffer from an overestimation bias. To mitigate this, various strategies have been developed, such as Double Q-learning (Hasselt, 2010) which aims to reduce overestimation by maintaining two separate estimators and using them to cross-validate each other's updates. This approach effectively dampens the positive bias introduced during the value estimation process. Similarly, methods like Weighted Q-learning (D'Eramo et al., 2016) have sought to refine the estimation of value functions by incorporating uncertainty measures.

Speedy Q-learning (Azar et al., 2011; John et al., 2020) seeks to accelerate the convergence towards optimal value estimates by adjusting updates towards more stable targets, indirectly addressing the overestimation issue by providing a more conservative update mechanism. RQ Learning (Tateo et al., 2017), another variant, adjusts the learning rate dynamically based on the uncertainty of the estimation, offering a nuanced approach to balancing exploration with the exploitation of known information.

A separate class of solutions involves model-based approaches. RMAX Brafman & Tennenholtz (2002) and OIM(Szita & Lőrincz, 2008) make subtly different choices that promote "optimism under uncertainty" to encourage exploration. MORMAX (Szita & Szepesvári, 2010) introduced a more nuanced mechanism for updating the model. Whereas, MBIE (Strehl & Littman, 2008) enhances exploration by maintaining confidence intervals for transition dynamics and rewards.

1.2 Tabular Reinforcement learning

Reinforcement learning problems are classically formalised as Markov Decision Processes (MDPs) which are tuples of the form $M = (\mathcal{S}, \mathcal{A}(s), p_{S'}, p_R)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}(s)$ the set of actions available at state $s \in \mathcal{S}$, $p_{S'}$ is a conditional probability distribution of the form $p_{S'}(\cdot \mid s, a)$ which characterises the random state S' reached after taking action $a \in \mathcal{A}(s)$ in state $s \in \mathcal{S}$, and p_R is a conditional probability distribution of the form $p_R(\cdot \mid s, a, s')$ that characterises the random reward R given the state transitioned from s to $s' \in \mathcal{S}$ under action a. Here we will assume the reward and transition distributions belong to the parametric families $\{p_R(\cdot \mid s, a, s', \psi) : \psi \in \Psi\}$ and $\{p_{S'}(\cdot \mid s, a, \theta) : \theta \in \Theta\}$, respectively, and consequently can be completely described by parameters $\psi \in \Psi$ and $\theta \in \Theta$. We will denote by $\psi^\star$ and $\theta^\star$ the true value of these parameters. At every time step, t, the agent will observe a state s_t and choose an action $a_t \in \mathcal{A}(s_t)$. The agent will then observe the next state s_{t+1} according to the probability distribution $p_{S'}(\cdot \mid s, a, \theta^\star)$ and receive a reward drawn from the distribution $p_R(\cdot \mid s, a, s', \psi^\star)$.

A policy is a mapping from each given state s to a probability distribution over the set $\mathcal{A}(s)$ of actions at s. The agents goal is to find a policy, π, that will maximise the cumulative expected discounted reward which is given by $\mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{k+1}]$, where γ is the discount factor which controls the value placed on future rewards and R_{k+1} is the reward received after the k-th action.

1.3 Possibility Theory

Possibility theory offers an alternative to probability theory when characterising uncertainty by specifically modelling epistemic uncertainty. This differentiation reflects the fundamentally different aims when modelling each type of uncertainty. In scenarios where unknown yet fixed quantities are involved, like model parameters, our uncertainty originates from an information deficit rather than randomness. Therefore, the focus shifts from employing probability distributions to quantify potential parameter values to assessing the credibility of these values by effectively modeling our level of information.

Credibilities describe how compatible an event is with our current information and take values between zero and one. Consider modeling the uncertainty about the true value $\psi^\star$ of a fixed but unknown parameter in a set Ψ. For example, for a given event E defined as the event $\psi^\star \in A$ for some subset A of Ψ, a credibility of one, $\text{cred}(E) = 1$, indicates total plausibility of the event, signifying no contradicting information against E. Conversely, a credibility of zero, $\text{cred}(E) = 0$, is obtained when the available information conclusively negates the event. Credibilities are monotonic, meaning that if available information suggest one event, E, is more plausible than another, E', the former should be assigned a credibility at least as high as the latter, $\text{cred}(E) \geq \text{cred}(E')$. Consequently, two equally plausible events will be assigned equal credibilities. If there is no objection against the event E, that is $\text{cred}(E) = 1$, then there cannot be any objection against the event E or E', which shows that $\text{cred}(\cdot)$ cannot be additive, and we consider $\text{cred}(E \text{ or } E') = \max\{\text{cred}(E), \text{cred}(E')\}$, which satisfies the above requirement. In particular, for an event E of the form $\psi^\star \in A$, it exists a function f such that $\text{cred}(\psi^\star \in A) = \sup_{\psi \in A} f(\psi)$, with f being a non-negative function on Ψ with supremum equal to 1, which we refer to as a possibility function.

In this paper we will apply the framework of Houssineau (2018) that combines probability theory and classical possibility theory into a unified model for both types of uncertainty. To illustrate this approach consider the toy example of a multi-armed bandit with just a single arm. This is still a MDP but with a single state and a single action, which are therefore omitted from notations. The agent draws the single arm repeatedly and receives rewards which are described by a random variable R distributed according to $p(\cdot \,|\, \psi^\star)$. The agent knows the parametric family to which $p(\cdot \,|\, \psi^\star)$ belongs so it will gain information on the value of $\psi^\star$ as it repeatedly draws the single arm. The aim of our approach is to model the uncertainty on $\psi^\star$ using possibility theory.

To this end we introduce an analogue of the notion of random variable in probability theory. Let Ω be a sample space, which can be interpreted as containing all possible states of nature. Note that in possibility theory there is no probability distribution associated with the sample space. Instead there is one true outcome, $\omega^\star \in \Omega$, interpreted as the "true state of nature". To formalise the relationship between the sample space and parameter space we define a function, $\boldsymbol{\psi} : \Omega \to \Psi$. Such functions are referred to as deterministic uncertain variables (d.u.v.s) and satisfy $\boldsymbol{\psi}(\omega^\star) = \psi^\star$ by construction.

To describe our current level of knowledge about $\psi^\star$ we use a possibility function $f_{\boldsymbol{\psi}} : \Psi \to [0, 1]$ (analogous to a probability density function), where $f_{\boldsymbol{\psi}}(\psi)$ is the credibility of $\boldsymbol{\psi} = \psi$, or analogously the credibility of $\psi^\star = \psi$. As there is no underlying probability distribution associated to the sample space there are no restrictions on the shape of $f_{\boldsymbol{\psi}}$ other than the normalisation $\sup_{\psi \in \Psi} f_{\boldsymbol{\psi}}(\psi) = 1$. To emphasise that the fact that a d.u.v. does not characterise the associated possibility function, we say that $f_{\boldsymbol{\psi}}$ *describes* $\boldsymbol{\psi}$. Indeed, the relationship between $f_{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}$ depends on the current level of information, so there is not a unique possibility function that describes $\boldsymbol{\psi}$. In the case of perfect information we have that $f_{\boldsymbol{\psi}} = \mathbf{1}_{\{\psi^\star\}}$ and if we have no information $f_{\boldsymbol{\psi}} \equiv 1$. It is notable that this makes defining uninformative priors trivial in all contexts, while by comparison this can be very challenging with probability theory.

We consider two notions of expectation for uncertain variables as $\mathbb{E}^*(\boldsymbol{\psi}) = \arg\max_{\psi \in \Psi} f_{\boldsymbol{\psi}}(\psi)$ and

$$\overline{\mathbb{E}}(\varphi(\boldsymbol{\psi})) = \sup_{\psi \in \Psi} \varphi(\psi) f_{\boldsymbol{\psi}}(\psi), \qquad \varphi : \Psi \to [0, \infty). \tag{1}$$

The expected value $\mathbb{E}^*(\boldsymbol{\psi})$ corresponds to the most plausible potential value for $\psi^\star$ given the information contained within $f_{\boldsymbol{\psi}}$. Throughout this paper we will assume $\mathbb{E}^*(\boldsymbol{\psi})$ is a singleton. This fits with notion of a single parameter value, $\psi^\star$. This notion does not necessarily correlate to the true parameter value at all times but instead to the most plausible one given the current information.

The second notion of expectation defined in (1) can be thought of as a maximum expected value. Given the current information level it balances the function value, $\varphi(\psi)$, with the possibility of the parameter value, $f_{\boldsymbol{\psi}}(\psi)$. This in-built balance between uncertainty and value mirrors that of the exploration-exploitation dilemma and is key to the approach we take within this paper.

As in the case of probability density functions, possibility functions are simply a more convenient way of expressing an underlying set function $\overline{\mathbb{P}}_{\boldsymbol{\psi}}(A) \doteq \text{cred}(\boldsymbol{\psi} \in A) = \sup_{\psi \in A} f_{\boldsymbol{\psi}}(\psi)$, $A \subseteq \Psi$. Recalling the earlier discussion of credibility we can see that $\overline{\mathbb{P}}_{\boldsymbol{\psi}}(A)$, which we refer to as the possibility of A, meets the criteria for describing epistemic uncertainty and hence possibility and credibility are used interchangeably. Such functions will satisfy the conditions of an outer measure with the additional property that $\overline{\mathbb{P}}_{\boldsymbol{\psi}}(\Psi) = 1$ and are called outer probability measures (o.p.m.s).

To incorporate both the random and deterministic uncertainty into a single model we need to consider more general o.p.m.s than $\overline{\mathbb{P}}_{\boldsymbol{\psi}}$. As before, let the random variable, R, describes the reward and be distributed according to $p_R(\cdot \,|\, \psi^\star)$. Assuming we do not know the true value of $\psi^\star$ we can consider the credibility of an event $A \times B$ where A is some subset of Ψ and B a measurable subset of $\mathbb{R}$. We have the joint o.p.m.

$$\overline{\mathbb{P}}_{\boldsymbol{\psi},R}(A \times B) = \sup_{\psi \in A} f_{\boldsymbol{\psi}}(\psi) \int_B p_R(r \,|\, \psi) dr,$$

where $f_{\boldsymbol{\psi}}$ describes $\boldsymbol{\psi}$ with our current level of knowledge. As opposed to joint probability distributions, the values taken by an o.p.m. such as $\overline{\mathbb{P}}_{\boldsymbol{\psi},R}$ depend on the order between the variables since the integral and supremum cannot be interchanged in general.

We can apply Bayes' rules to this expression, following the methods in Houssineau (2018), to obtain

$$f_{\boldsymbol{\psi}}(\psi \,|\, r) = \frac{p_R(r \,|\, \psi) f_{\boldsymbol{\psi}}(\psi)}{\sup_{\phi \in \Psi} p_R(r \,|\, \phi) f_{\boldsymbol{\psi}}(\phi)}, \tag{2}$$

where r is a realisation of R, which is similar to Bayes' rule in probability theory with the key difference being the supremum, rather than the integral, in the denominator. This normalisation criteria is often simpler to evaluate or approximate numerically when compared to the usual integral. This leads to posterior functions being often easier to compute than in the standard probabilistic context, while still enjoying asymptotic guarantees (Houssineau et al., 2019). The considered analogue of Bayesian inference has been used successfully with complex hidden Markov models (Houssineau & Bishop, 2018; Houssineau, 2021).

In the explicit examples given in Section 4 we will make use of rewards distributed according to Bernoulli random variables. Using Bayes' rule above, after observing realisations of R and obtaining α successes and β failures. We will be obtain the conjugate prior for the Bernoulli probability distribution which we call the beta possibility function,

$$\overline{\text{Be}}(\theta; \alpha, \beta) = \frac{\wp(\alpha + \beta)}{\wp(\alpha)\wp(\beta)} \theta^\alpha (1 - \theta)^\beta, \tag{3}$$

where $\wp(x) = x^x$. The parameters are shifted from the standard beta density function. This allows us to interpret the case $\alpha = \beta = 0$ as the uninformative possibility function.

2 Markov Reward Process

This section considers the simplified setting of a Markov Reward Process (MRP) to illustrate how possibility theory can be used to model the reward and transition parameters without the added

complexity of incorporating actions. MRPs can be seen as a special case of MDPs where the the action set $\mathcal{A}(s)$ is a singleton for all $s \in \mathcal{S}$ and there is therefore no need to devise a policy; instead transitions from a state, s, follow the distribution $p_{S'}(\cdot \,|\, s, \theta)$ for some parameter $\theta \in \Theta$, where the actions are omitted from the notation. We denote by $N(s) \subseteq \mathcal{S}$ the set of neighbouring states of s, that is the set of points s' such that $p_{S'}(s' \,|\, s, \theta) > 0$ for some $\theta \in \Theta$.

Considering the undiscounted setting, as is standard for MRPs, the true value $v^\star(s)$ at $s \in \mathcal{S}$ is

$$v^\star(s) = \mathbb{E}\big[\mathbb{E}[R \,|\, s, S', \psi^\star] + v^\star(S') \,|\, \theta^\star\big] \doteq B_s(v^\star\restriction_{N(s)}, \psi^\star, \theta^\star), \tag{4}$$

where $v^\star\restriction_{N(s)}$ is the collection of values $(v^\star(s'))_{s' \in N(s)}$, that is $v^\star$ restricted to $N(s)$. However, since we do not know $\theta^\star$, $\psi^\star$ and $v^\star\restriction_{N(s)}$, we model them as d.u.v.s, which turns the value function into a d.u.v. itself, characterised by $\boldsymbol{v}(s) = B_s(\boldsymbol{v}\restriction_{N(s)}, \boldsymbol{\psi}, \boldsymbol{\theta})$, with B_s defined in (4).

One could characterise the possibility function describing $\boldsymbol{v}(s)$ given some information about $\boldsymbol{v}(s')$, $s' \in N(s)$, and about $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, under the form of possibility functions $f_{\boldsymbol{v}\restriction_{N(s)}}$, $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\theta}}$, respectively. For instance, one could assume that the reward parameter $\boldsymbol{\psi}$ and the transition parameter $\boldsymbol{\theta}$ are independent, i.e., $f_{\boldsymbol{\psi},\boldsymbol{\theta}}(\psi, \theta) = f_{\boldsymbol{\psi}}(\psi) f_{\boldsymbol{\theta}}(\theta)$ for any $(\psi, \theta) \in \Psi, \Theta$, and that $f_{\boldsymbol{v}\restriction_{N(s)}}$ is the indicator of the collection $v^*_{\psi,\theta}\restriction_{N(s)} = (v^*_{\psi,\theta}(s'))_{s' \in N(s)}$, i.e., $\boldsymbol{v}(s')$ is known to be equal to $v^*_{\psi,\theta}(s')$ for any $s' \in N(s)$ when the reward parameter ψ and the transition parameter θ are fixed. In particular, it holds that $v^\star(s) = v^*_{\psi^\star,\theta^\star}(s)$ for all $s \in \mathcal{S}$. It would then hold that

$$f_{\boldsymbol{v}(s)}(v) = \sup\big\{f_{\boldsymbol{\psi}}(\psi) f_{\boldsymbol{\theta}}(\theta) : \psi \in \Psi, \theta \in \Theta, v = B^*_{s,N(s)}(\psi, \theta)\big\}, \tag{5}$$

where $B^*_{s,N(s)} : (\psi, \theta) \mapsto B_s(v^*_{\psi,\theta}\restriction_{N(s)}, \psi, \theta)$, which is simply the change of variable formula for possibility functions corresponding to (4). We can check that if there is no information about any of the quantities of interest, that is $f_{\boldsymbol{\psi}} \equiv 1$ and $f_{\boldsymbol{\theta}} \equiv 1$, then $f_{\boldsymbol{v}(s)}$ is the indicator of the image of $B^*_{s,N(s)}$, that is, there is no objection against any of the values at s that are allowed by the model. However, even when $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\theta}}$ take simple parametric forms, the non-linearity of $B^*_{s,N(s)}$ means that $f_{\boldsymbol{v}(s)}$ can be challenging to characterise and/or approximate efficiently. In addition, the assumption that the value function is known on $N(s)$ given θ and ψ is unrealistic and does not correspond to the way the information about the value flows between different states. This is particularly obvious in settings where it is possible to transition from s to itself, that is where $s \in N(s)$. Possibility theory allows modelling of this type of uncertainty: despite the fact that a quantity *could* be known, in this case the value of states in $N(s)$ given ψ and θ, one can model the uncertainty due to the difficulty in computing the said quantity. This is well defined thanks to the non-uniqueness of possibility functions; indeed, possibility functions model the information *available* to the modeller, which can vary from one modeller to the other. This highlights the subjectivity inherent to possibility functions, which is unavoidable when dealing with information rather than randomness. Therefore, one can model the current available information about $\boldsymbol{v}\restriction_{N(s)}$ through a non-degenerate possibility function $f_{\boldsymbol{v}\restriction_{N(s)}}(\cdot \,|\, \psi, \theta)$, that is a possibility function whose support is not a singleton.

Given a dataset $\mathcal{D}_t = \{(s_{t'}, r_{t'}, s'_{t'})\}_{t' \leq t}$ with $s_{t'}$, $r_{t'}$ and $s'_{t'}$ the state, reward and next state at a given epoch $t' \in \{1, \dots, t\}$, we assume that $\boldsymbol{v}$ is conditionally independent of $\mathcal{D}$ given $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$, i.e., $f_{\boldsymbol{v}}(\cdot \,|\, \psi, \theta, \mathcal{D}) = f_{\boldsymbol{v}}(\cdot \,|\, \psi, \theta)$. This is a meaningful assumption since the true value function $v^\star$ of $\boldsymbol{v}$ could be known when $\psi^\star$ and $\theta^\star$ are known, without any additional information. It follows from this and from the independence of $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ that $f_{\boldsymbol{v},\boldsymbol{\psi},\boldsymbol{\theta}}(v, \psi, \theta \,|\, \mathcal{D}) = f_{\boldsymbol{v}}(v) f_{\boldsymbol{\psi}}(\psi \,|\, \mathcal{D}) f_{\boldsymbol{\theta}}(\theta \,|\, \mathcal{D})$, so that we can simply update the information about $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ separately whenever new data are available, e.g.

$$f_{\boldsymbol{\psi}}(\psi \,|\, \mathcal{D}_{t+1}) = \frac{p_R(r_{t+1} \,|\, s_{t+1}, s'_{t+1}, \psi) f_{\boldsymbol{\psi}}(\psi \,|\, \mathcal{D}_t)}{\sup_{\psi' \in \Psi} p_R(r_{t+1} \,|\, s_{t+1}, s'_{t+1}, \psi') f_{\boldsymbol{\psi}}(\psi' \,|\, \mathcal{D}_t)},$$

which allows us to leverage the possibilistic analogue of conjugate prior families.

It remains that $f_{\boldsymbol{v}(s)}$ will be challenging to characterise in general, and instead, we consider potential estimators of $\boldsymbol{v}(s)$ based on the two considered notions of expected value:

$$\mathbb{E}^*(\boldsymbol{v}(s)) = \mathbb{E}^*\left(B_s(\boldsymbol{v}\restriction_{N(s)}, \boldsymbol{\psi}, \boldsymbol{\theta})\right) = B_s\big(\mathbb{E}^*(\boldsymbol{v}\restriction_{N(s)}), \mathbb{E}^*(\boldsymbol{\psi}), \mathbb{E}^*(\boldsymbol{\theta})\big) \tag{6}$$

and

$$\overline{\mathbb{E}}(\boldsymbol{v}(s)) = \overline{\mathbb{E}}\left(B_s(\boldsymbol{v}\restriction_{N(s)}, \boldsymbol{\psi}, \boldsymbol{\theta})\right) \tag{7a}$$

$$= \overline{\mathbb{E}}\left(\mathbb{E}\big[\mathbb{E}[R \,|\, s, S', \boldsymbol{\psi}] + \boldsymbol{v}(S') \,|\, \boldsymbol{\theta}\big]\right) \tag{7b}$$

$$= \overline{\mathbb{E}}\Big(\mathbb{E}\big[\bar{r}(s, S') + \boldsymbol{v}(S') \,|\, \boldsymbol{\theta}\big]\Big), \tag{7c}$$

where $\bar{r}(s, s') = \overline{\mathbb{E}}(\mathbb{E}[R \,|\, s, s', \boldsymbol{\psi}])$ is assumed to be non-negative, and where we have assumed that the independence of the information about the reward when reaching different states $s' \in N(s)$. Indeed, the sum defining $\mathbb{E}(\cdot)$ and the maximum defining $\overline{\mathbb{E}}(\cdot)$ can be interchanged if each term in the sum is maximised independently. The assumption that $\bar{r}(s, s') \geq 0$ for any $s, s' \in \mathcal{S}$ can be satisfied in most cases by shifting the rewards, a strategy commonly used in deep RL, see e.g. Sun et al. (2022), albeit for different reasons. In our experimental results, we adopt the alternative approach of shifting rewards solely when calculating the optimal parameters of the maximum-expected value. The optimal parameters can then be used with the original rewards to calculate the optimistic estimate of the true rewards. That is, we first compute $\theta^* = \arg\max_\theta f(\theta)(\varphi(\theta) + c)$ where c ensures that $\varphi(\theta) + c \geq 0$ and then define the maximum expected value as $f(\theta^*)\varphi(\theta^*)$.

The expected value $\mathbb{E}^*(\boldsymbol{v}(s))$ corresponds to the most credible value for the state given our current information, which is equal to the deterministic combination of the most credible values for the d.u.v.s defining it, as is intuitive for deterministic quantities. This is in contrast to the probabilistic notion of expected value, which does not transform coherently under nonlinear mappings, as should be the case for random quantities. The expected value $\mathbb{E}^*(\boldsymbol{v}(s))$ can be set-valued, e.g., when the possibility functions describing $f_{\boldsymbol{v}(s')}$, $f_{\boldsymbol{\psi}}$ and $f_{\boldsymbol{\theta}}$ are uninformative. In any case, the expected value $\mathbb{E}^*(\boldsymbol{v}(s))$ does not balance exploration and exploitation since it completely ignores the uncertainty modelled by $f_{\boldsymbol{v}(s)}$, except when it is set-valued, and would therefore lead to an exploitation-based policy. This is in contrast to the alternative notion of expected value $\overline{\mathbb{E}}(\boldsymbol{v}(s))$, which provides an optimistic estimate of the value of the state s, by maximising the values at s' and the underlying expected reward and transitions against their respective possibilities.

From a practical viewpoint, the expression of the expected value $\overline{\mathbb{E}}(\boldsymbol{v}(s))$ given in (7c) is not satisfactory as it requires defining a possibility function describing $\boldsymbol{v}\restriction_{N(s)}$. To further simplify the recursion, we consider instead an upper bound of the maximum expected value defined as $\bar{v}(s) = \overline{\mathbb{E}}_{S'}\big[\bar{r}(s, S') + \overline{\mathbb{E}}_{\boldsymbol{v}}(\boldsymbol{v}(S'))\big] \geq \overline{\mathbb{E}}(\boldsymbol{v}(s))$, where $\overline{\mathbb{E}}_{S'}[\cdot] = \overline{\mathbb{E}}_{\boldsymbol{\theta}}\big(\mathbb{E}[\cdot \,|\, \boldsymbol{\theta}]\big)$, with the inequality becoming an equality when $\boldsymbol{\psi}$ and $\boldsymbol{v}(s')$, $s' \in N(s)$, are mutually independent. Although this assumption does not hold in general, this inequality hints at the potential of defining a closed-form recursion based on $\bar{v}(s)$: plugging in $\bar{v}(S')$ instead of $\overline{\mathbb{E}}_{\boldsymbol{v}}(\boldsymbol{v}(S'))$, we obtain the recursion

$$\bar{v}(s) = \overline{\mathbb{E}}_{S'}\big[\bar{r}(s, S') + \bar{v}(s)\big],$$

where the value in the left and right hand sides are represented by the same symbol, following a standard abuse of notations in this context.

2.1 GridWorld Experiment

Consider the set $\mathcal{S}$ to be a 2-dimensional 5x5 GridWorld. The 25 states are labelled as (n, m) for $n, m \in \{0, \ldots, 4\}$ and the available transitions at each state are moving up, right, down and left, unless the agent is at the edge then a transition that would take it outside of the 5×5 grid will have no effect. Rewards for transitioning from one state to the next are Bernoulli with probability of success $p_R(1 \,|\, s, s', \psi) = \psi_{s,s'}$, i.e. the parameter ψ is taken to be the 25×25 matrix of unknown probabilities of success. We assume that each $\boldsymbol{\psi}_{s,s'}$ is described by a beta possibility function $\overline{\mathrm{Be}}(\alpha_{s,s'}, \beta_{s,s'})$. Similarly, we assume that the transition probabilities are parameterised as $p_{S'}(s' \,|\, s, \theta) = \theta_{s,s'}$, with $\boldsymbol{\theta}_s \doteq (\boldsymbol{\theta}_{s,s'})_{s' \in N(s)}$ being described by a Dirichlet possibility function, defined as

$$f_{\boldsymbol{\theta}_s}(\theta_s) = \overline{\mathrm{Dir}}(\theta_s; \alpha_s) \doteq \wp\Big(\sum_{s' \in N(s)} \alpha_{s,s'}\Big) \prod_{s' \in N(s)} \frac{\theta_{s,s'}^{\alpha_{s,s'}}}{\wp(\alpha_{s,s'})},$$

for some collection of parameters $\alpha_s = (\alpha_{s,s'})_{s' \in N(s)}$. The uncertain value function can now be simplified to $\boldsymbol{v}(s) = \sum_{s' \in N(s)} (\boldsymbol{\psi}_{s,s'} + \boldsymbol{v}(s'))\boldsymbol{\theta}_{s,s'}$, where we have used the fact that $\mathbb{E}[R \,|\, s, s', \boldsymbol{\psi}] = \boldsymbol{\psi}_{s,s'}$. It follows that

$$\bar{v}(s) = \overline{\mathbb{E}}\left(\sum_{s' \in N(s)} \big(\overline{\mathbb{E}}(\boldsymbol{\psi}_{s,s'}) + \bar{v}(s')\big)\boldsymbol{\theta}_{s,s'} \right), \tag{8}$$

where $\overline{\mathbb{E}}(\boldsymbol{\psi}_{s,s'}) = \wp(\alpha_{s,s'} + \beta_{s,s'})\wp(\alpha_{s,s'} + 1)/(\wp(\alpha_{s,s'})\wp(\alpha_{s,s'} + \beta_{s,s'} + 1))$, so we are left with the optimisation problem

$$\max_{\theta_s \in \Theta_s} \overline{\mathrm{Dir}}(\theta_s; \alpha_s) \sum_{s' \in N(s)} \big(\overline{\mathbb{E}}(\boldsymbol{\psi}_{s,s'}) + \bar{v}(s')\big)\, \theta_{s,s'}, \tag{9}$$

where Θ_s is the $|N(s)| - 1$ standard simplex indexed by $N(s)$, that is $\Theta_s = \big\{(\theta_{s,s'})_{s' \in N(s)} : \sum_{s' \in N(s)} \theta_{s,s'} = 1, \theta_{s,s'} \in [0, \infty), s' \in N(s)\big\}$. This setting is simple enough that the true value for each state can be computed explicitly.

The experimental results displayed in the subplot titled MRP in Figure 1 correspond to the case where: i) the agent starts each episode at $s = (2, 2)$, in the centre of the grid, ii) there are two terminal states at the the top-left $(0, 0)$ and bottom right $(4, 4)$, iii) it holds that $\psi^\star_{s,a,s'} = 0.5$ for any s, s', except when s' is terminal, in which case $\psi^\star_{s,s'} = 1$, and iv) it holds that $\theta^\star_{s,s'} = (1, \ldots, 1)/|N(s)| \in [0, 1]^{|N(s)|}$. Note that results are aggregated across 100 experimental runs.

Leveraging the proposed model to increase the accuracy of the value estimates, the approaches of (6) and (8) outperform TD(0), as was expected due to the model-based elements in our approach. More notably, the optimistic estimates of (8) quickly coincide with those of the most credible model, demonstrating our approach's ability to discard optimism as information is gained. MRPs are purely an estimation problem, which is why (6) can perform well. We will focus on (8) when generalising to MPDs since the maximum expected value has the potential to help balancing exploration and exploitation.

3 Proposed Approach

3.1 Maximum expected Q-values

Based on the insights obtained in Section 2, we can now generalise the proposed methodology to full MDPs, i.e., MDPs with a non-trivial action set $\mathcal{A}(\cdot)$. In order to keep the discussion as general as possible, we introduce $\tilde{\mathbb{E}}(\cdot \,|\, s)$ as the expectation operator over the action set $\mathcal{A}(s)$, for any $s \in \mathcal{S}$, without defining it more specifically for now. The uncertain value $\boldsymbol{Q}(s, a)$ at the state action pair (s, a) with $a \in \mathcal{A}(s)$ can now be defined as

$$\boldsymbol{Q}(s, a) = \mathbb{E}\Big[\mathbb{E}[R \,|\, s, a, S', \boldsymbol{\psi}] + \tilde{\mathbb{E}}\big(\boldsymbol{Q}(S', a') \,|\, S'\big) \,|\, \boldsymbol{\theta}\Big]. \tag{10}$$

Three models for the next action a' can be considered:

- M.1 if a' is a random action A' from the current policy π then $\tilde{\mathbb{E}}(\cdot \,|\, s)$ is the standard expected value and (10) is related to the expected SARSA algorithm.
- M.2 if a' is an uncertain action $\boldsymbol{a}'$ described by a possibility function $f_{\boldsymbol{a}'}$ and $\tilde{\mathbb{E}}(\cdot \,|\, s)$ is defined based on the notion of expected value $\overline{\mathbb{E}}(\cdot)$, with $f_{\boldsymbol{a}'}$ unrelated to the considered policy, then the recursion can be considered off-policy. If there is no available information about $\boldsymbol{a}'$, so that $f_{\boldsymbol{a}'} \equiv 1$, then (10) is related to Q-learning.
- M.3 if a' is an uncertain action $\boldsymbol{a}'$ described by a possibilistic policy $\bar{\pi}$ and $\tilde{\mathbb{E}}(\cdot \,|\, s)$ is defined based on the notion of expected value $\mathbb{E}^*(\cdot)$, then (10) is related to to SARSA with a greedy policy.

Based on the advantages of off-policy algorithms and the success of Q-learning in practice, we consider M.2 with $f_{\boldsymbol{a}'} \equiv 1$ in our experiments.

Following the same approach and assumptions as in Section 2, we can obtain a closed-form estimation for the maximum expected value of a state-action pair (s, a) as

$$\bar{Q}(s,a) = \overline{\mathbb{E}}_{S'}\Big[\bar{r}(s,a,S') + \tilde{\mathbb{E}}\big(\bar{Q}(S',a') \,|\, S'\big)\Big], \tag{11}$$

where we have extended the assumptions of independence between values $\boldsymbol{v}(s)$ at different states s to Q-values $\boldsymbol{Q}(s,a)$ at different state-action pairs (s,a).

Initialisation The estimates $\bar{Q}(s,a)$ obtained with the proposed approach are upper bounds for the maximum expected value $\overline{\mathbb{E}}(\boldsymbol{Q}(s,a))$. Therefore, it would be natural to initialise at an upper bound for the true Q-value $Q^\star(s,a)$ when there is no additional information. The existence of this upper bound depends on the environment and its use in the prior knowledge that the agent has access to. In a setting where there is a known maximum reward, $r_{\max}$, and discount γ we can use the estimate $r_{\max}/(1-\gamma)$. Additional information, such as knowledge of the number states can be useful to further refine this estimate. However, overestimating the initialisation value is insignificant as the algorithm quickly forgets the initial value in exchange for the information it has gathered. In practice this means any upper bound on Q-values can be selected. When implementing the initialisation value care should be taken to ensure that the value of terminal states is set to 0. This can be done either during initialisation if the terminal states are known, or when they are first encountered through interaction.

With the proposed initialisation, the recursion defined in (11) behaves as follows. Unexplored state-actions pairs will tend to have a relatively large value, comparable or larger than the optimal value. This means that there will be a natural tendency towards exploration until enough information is gathered about non-optimal trajectories, at which point exploitation will become dominant. The main consequence of this behaviour is that most of the exploration-exploitation trade-off is achieved at the level of the Q-values.

3.2 Design of the policy

As hinted at above, there are multiple ways to define a policy, either as a probability distribution π or as a possibility function $\bar{\pi}$ over the relevant set of actions. As is usual, we are interested in balancing exploration and exploitation. Since randomness is often used as the basis for exploration, it is useful to construct a probabilistic policy π from which actions can be sampled, even when π is deduced from a possibilistic policy $\bar{\pi}$. This is relevant in situations where $\bar{\pi}$ can be defined in a principled way.

We will consider two ways of defining a probability distribution π based on a possibility function $\bar{\pi}$. The first one simply defines π via $\pi(a \,|\, s) \propto \bar{\pi}(a \,|\, s)$ for all $a \in \mathcal{A}(s)$, which is always possible when $\mathcal{A}(s)$ is finite. This is a pragmatic approach that simply interprets the possibility $\bar{\pi}(a \,|\, s)$ as an odds ratio against the most likely action. Indeed, it holds that

$$\frac{\pi(a \,|\, s)}{\pi(a^*_{\bar{\pi}}(s) \,|\, s)} = \frac{\bar{\pi}(a \,|\, s)}{\bar{\pi}(a^*_{\bar{\pi}}(s) \,|\, s)} = \bar{\pi}(a \,|\, s)$$

with $a^*_{\bar{\pi}}(s) = \arg\max_{a \in \mathcal{A}(s)} \bar{\pi}(a \,|\, s)$. The second approach is based on the interpretation of o.p.m.s. as upper bounds for probability distributions. In particular, we might want to select a policy belonging to $\Pi = \{\pi : \sum_{a \in A} \pi(a \,|\, s) \leq \max_{a \in A} \bar{\pi}(a \,|\, s)\}$, so as to ensure that the possibility $\max_{a \in A} \bar{\pi}(a \,|\, s)$ is indeed an upper bound for the probability of the event $a \in A$. To select a specific policy in Π, we might want to maximise the exploration whenever there is uncertainty, so that one natural option is to choose the distribution $\pi^* \in \Pi$ with the maximum entropy.

It remains to define the policy π or $\bar{\pi}$ more specifically. The standard ϵ-greedy policy can easily be defined based on the action $a^*(s) = \arg\max_{a \in \mathcal{A}(s)} \bar{Q}(s,a)$. However, we are interested in policies that further leverage the properties of $\boldsymbol{Q}(s,a)$ to balance exploration and exploitation, instead of introducing a hyperparameter such as ϵ. If we had access to the possibility function $f_{\boldsymbol{Q}}$, we could

define $\bar{\pi}$ based on the possibility that a given action has higher value then the greedy action, e.g., $\bar{\pi}(a \mid s) = \overline{\mathbb{P}}\big(\boldsymbol{Q}(s,a) > \boldsymbol{Q}(s,a^*(s))\big)$, which is the possibility of action a improving upon the greedy action. In the proposed recursion (11), the possibility function $f_{\boldsymbol{Q}}$ is not available. Yet, as noted above, the degree of exploration is naturally balanced by the maximum expected Q-values so that a simple policy is sufficient. To illustrate this and to avoid introducing tuning parameters, we simply consider the greedy policy $\pi(\cdot \mid s) = \delta_{a^*(s)}$ in our experiments. We name this approach Possibilistic Q-learning, or PQL; the simplified algorithm is presented in Algorithm 1.

Algorithm 1 PQL

Initialize Q, History, s
repeat
  Choose $a \leftarrow \arg\max_a \bar{Q}(s,\cdot)$, observe r, s'
  Update History with (s, a, r, s')
  $\bar{Q}(s,a) \leftarrow \overline{\mathbb{E}}_{S'}\left[\bar{r}(s,a,S') + \arg\max_{a'} \bar{Q}(S',a')\right]$
  $s \leftarrow s'$
until end

4 Experimental Results

In this section we benchmark the PQL algorithm across six GridWorld environments implemented in Mushroom RL (D'Eramo et al., 2021). Results are aggregated across 100 experimental runs per algorithm, with ten evaluation episodes run after each training episode to account for stochasticity. Seven baseline algorithms were also trained on each environment: Q-Learning, Speedy Q-Learning, SARSA, SARSA(0.9), RQ-Learning, Delayed Q-Learning, and R-Max. Hyperparameters for the baseline approaches were manually optimised for each environment. Note that the PQL algorithm does not depend on hyperparameters.

We consider the following tasks:

Standard As a basic test we deploy the algorithms on a 7x7 grid, episodes begin at (0,0) with a single absorbing state at (6,6) where a reward of 10 is received. A discount factor of 0.9 is used.

Standard with Penalty To test the impact of the reward shifting within our approach we repeat the standard environment but with penalty of -1 for every transition towards a non-absorbing state.

Windy Cliffworld To test the resilience to random noise on actions we consider a CliffWorld environment. This is a 5x4 GridWorld where episodes start at (1,1) with the goal of reaching the absorbing state (1,5) where a reward of 5 is received. (1,2), (1,3) and (1,4) are absorbing states where no reward is received. At every time step there is a risk of being up (towards the cliff) with probability 0.3, this overrides the selected action.

Multi-peak Environment To test the ability of algorithms to differentiate between two close-to-optimal policies, a 5x5 GridWorld is implemented with two absorbing states (1,5) and (5,5), where the agent receives reward 9 and 10 respectively. Episodes begin at (3,1), a reward of 1 is achieved with probability 0.5 at each time step.

Absorbing Barriers To test the ability of our approach to quickly identify and ignore significantly suboptimal actions we create a GridWorld where there is a gap between absorbing states that must be traversed to reach the goal state. This is a 5x5 GridWorld, episodes start at (1,1), (5,5) is an absorbing state with reward 10. (3,1), (3,2), (3,4) and (3,5) are absorbing states where no reward is received. A reward of 1 is received with probability 0.5 at any non-absorbing state.

The experimental results are shown in Figure 1. PQL can be seen to be competitive in all tasks, achieving a significantly faster convergence than the baseline algorithms for the more complex tasks

such as Windy Cliffworld and Absorbing Barriers. PQL is only outperformed on the relatively simple environments where the eligibility traces of SARSA(0.9) are more beneficial. However, even on these environments, PQL finds the true optimal solution, unlike SARSA(0.9). In more complex, stochastic environments, the benefits of eligibility traces diminish. Notably, except in the highly stochastic Windy CliffWorld, PQL consistently identifies and eventually adheres to an optimal policy. PQL goes through a period of exploration, after which, having sufficiently reduced uncertainty, it selects an optimal strategy. This behaviour emerges naturally from the inherent balance between exploration and exploitation provided by the maximum-expected value.

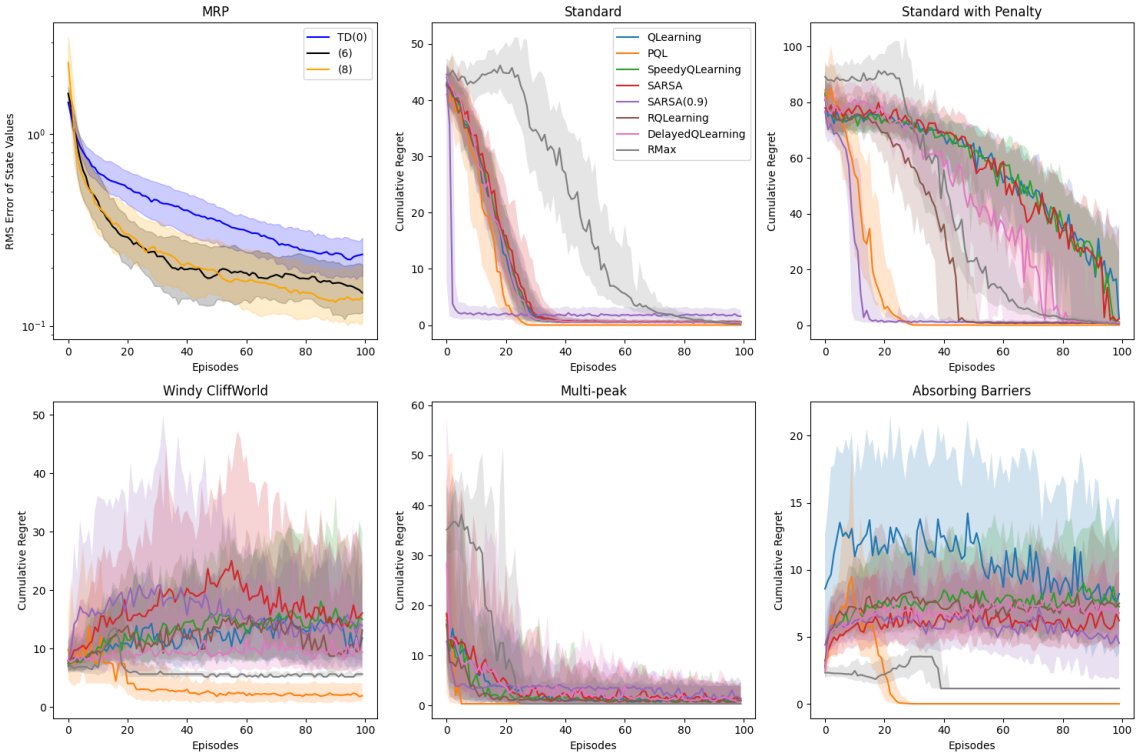

Figure 1: Performance assessment via the cumulative regret for MDPs and via the root mean square (RMS) error for the MRP. Lines are based on the median and shaded areas correspond to an interquartile range between the first and third quartiles, calculated across 100 realisations.

5 Conclusion

We have shown how the use of a dedicated representation of epistemic uncertainty allows for introducing algorithms that behave differently depending on the type of uncertainty that they face. In particular, we have focused on a simple algorithm, PQL, which leverages models of the reward and transition to provide a tuning-free method that displays a strong performance in a range of experiments.

Future work will focus on the extension of the proposed approach to more challenging tasks requiring the use of deep reinforcement learning methodologies, where ensembles and/or variance networks could be reinterpreted through the lens of possibility theory to provide a generalisation of the proposed methodology. In addition, using optimistic strategies has received renewed attention in the context of multi-agent reinforcement learning (Wei & Luke, 2016; Jiang & Lu, 2022), where the potential lack of information about the policy of other agents is an important source of epistemic uncertainty, hence highlighting the potential of a possibilistic modelling in this context.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Christopher Atkeson. Nonparametric model-based reinforcement learning. *Advances in neural information processing systems*, 10, 1997.

Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy q-learning. In *Advances in neural information processing systems*, 2011.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Carlo D'Eramo, Davide Tateo, Andrea Bonarini, Marcello Restelli, and Jan Peters. Mushroomrl: Simplifying reinforcement learning research. *Journal of Machine Learning Research*, 22(131):1–5, 2021.

Didier Dubois and Henry Prade. Possibility theory and its applications: Where do we stand? *Springer handbook of computational intelligence*, pp. 31–60, 2015.

Carlo D'Eramo, Marcello Restelli, and Alessandro Nuara. Estimating maximum expected value through gaussian approximation. In *International Conference on Machine Learning*, pp. 1032–1040. PMLR, 2016.

Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value expansion for efficient model-free reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

Jeremie Houssineau. Parameter estimation with a class of outer probability measures. *arXiv preprint arXiv:1801.00569*, 2018.

Jeremie Houssineau. A linear algorithm for multi-target tracking in the context of possibility theory. *IEEE Transactions on Signal Processing*, 69:2740–2751, 2021.

Jeremie Houssineau and Adrian N Bishop. Smoothing and filtering with a class of outer measures. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):845–866, 2018.

Jeremie Houssineau, Neil K Chada, and Emmanuel Delande. Elements of asymptotic theory with outer probability measures. *arXiv preprint arXiv:1908.04331*, 2019.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110:457–506, 2021.

Jiechuan Jiang and Zongqing Lu. I2q: A fully decentralized Q-learning algorithm. *Advances in Neural Information Processing Systems*, 35:20469–20481, 2022.

Indu John, Chandramouli Kamanchi, and Shalabh Bhatnagar. Generalized speedy Q-learning. *IEEE Control Systems Letters*, 4(3):524–529, 2020.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Thomas M Moerland, Joost Broekens, Aske Plaat, Catholijn M Jonker, et al. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1):1–118, 2023.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Hao Sun, Lei Han, Rui Yang, Xiaoteng Ma, Jian Guo, and Bolei Zhou. Exploit reward shifting in value-based deep-RL: Optimistic curiosity-based exploration and conservative exploitation via linear reward shaping. *Advances in Neural Information Processing Systems*, 35:37719–37734, 2022.

Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pp. 216–224. Elsevier, 1990.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 1998.

István Szita and András Lőrincz. The many faces of optimism: a unifying approach. In *Proceedings of the 25th international conference on Machine learning*, pp. 1048–1055, 2008.

István Szita and Csaba Szepesvári. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1031–1038, 2010.

Davide Tateo, Carlo D'Eramo, Alessandro Nuara, Marcello Restelli, and Andrea Bonarini. Exploiting structure and uncertainty of bellman updates in markov decision processes. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. IEEE, 2017.

Ermo Wei and Sean Luke. Lenient learning in independent-learner stochastic cooperative games. *The Journal of Machine Learning Research*, 17(1):2914–2955, 2016.

Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020.