Uncertainty Propagation for Quality Assurance in Reinforcement Learning

Daniel Schneegass, Steffen Udluft, and Thomas Martinetz *Senior Member, IEEE*

***Abstract*— In this paper we address the reliability of policies derived by Reinforcement Learning on a limited amount of observations. This can be done in a principled manner by taking into account the derived Q-function's uncertainty, which stems from the uncertainty of the estimators used for the MDP's transition probabilities and the reward function. We apply uncertainty propagation parallelly to the Bellman iteration and achieve confidence intervals for the Q-function. In a second step we change the Bellman operator as to achieve a policy guaranteeing the highest minimum performance with a given probability. We demonstrate the functionality of our method on artificial examples and show that, for an important problem class even an enhancement of the expected performance can be obtained. Finally we verify this observation on an application to gas turbine control.**

I. Introduction and Related Work

REINFORCEMENT LEARNING (RL) [1] aims to derive an optimal policy from observations acquired by the exploration of an uncertain environment. For a limited amount of observations the collected information might not be sufficient to fully determine the environment's properties. Assuming the environment to be a Markov decision process (MDP), it is in general only possible to create estimators for the MDP's transition probabilities and the reward function. As the true parameters remain uncertain, the derived policy, which is optimal w.r.t. the estimators is in general not optimal w.r.t. the real MDP and might even perform insufficiently. This is unacceptable in industrial environments, where quality assurance is of particular importance.

To overcome this problem, we incorporate the uncertainties of the estimators into the derived Q-function, which is utilised by many RL methods. In order to guarantee a minimal performance with a given probability, as a solution to quality assurance, we present an approach using statistical uncertainty propagation (UP) [2] on the Bellman iteration to obtain Q-functions together with their uncertainty. In a second step, we introduce a modified Bellman operator, jointly optimising the Q-function and minimising its uncertainty. This method leads to a policy, which is no more optimal in the conventional meaning, but maximises the guaranteed minimal performance and hence optimises the quality requirements. In this paper we apply the technique exemplarily on discretised MDPs and outline a possible generalisation to function approximation.

Daniel Schneegass and Steffen Udluft are with the Learning Systems Department of the Siemens AG, Corporate Technology, Information & Communications; email: daniel.schneegass.ext@siemens.com, steffen.udluft@siemens.com. Thomas Martinetz is director of the Institute for Neuro- and Bioinformatics of the University of Luebeck; email: martinetz@informatik.uni-luebeck.de.

There have already been several contributions to estimate generalisation, confidence, and performance bounds in RL. We consider the work of Bertsekas et.al. [3], who gave lower-bounds on the policy's performance by using policy iteration techniques, which were substantially improved by Munos et.al. [4]. Kearns et.al. [5] discussed error-bounds for a theoretical policy search algorithm based on trajectory trees. Capacity results on policy evaluation are given by Peshkin et.al. [6]. Antos et.al. [7] provided a broad capacity analysis of Bellman residual minimisation in batch RL. Incorporating prior knowledge about confidence and uncertainty directly into the approached policy were already applied in former work as well in the context of Bayesian Reinforcement Learning. We especially mention the work of Engel et.al., Gaussian Process Temporal Difference Learning (GPTD) [8], [9] and a similar approach by Rasmussen and Kuss [10]. They applied Gaussian Processes and hence a prior distribution over value functions in RL, which is updated to posteriors by observing samples from the MDP. Engel et.al. recently developed algorithms for Bayesian Policy Gradient RL [11] and Bayesian Actor-Critic RL [12] as further model-free approaches to Bayesian RL. Gaussian Processes have the advantage of inherently providing a measure of uncertainty.

In model-based approaches, however, one starts with a natural local measure of the uncertainty of the transition probabilities and rewards. One of the first concerning work in the context of RL is provided by Dearden et.al. [13], [14], who applied Q-learning in a Bayesian framework with an application to the exploration-exploitation-trade-off. Poupart et.al. present, in their recent work [15], an approach for efficient online learning and exploration in a Bayesian context, they ascribe Bayesian RL to POMDPs. Most related to our approach is the recent independent work by Delage and Mannor [16], who solved the percentile optimisation problem by convex optimisation, and applied it to the exploration-exploitation-trade-off. They suppose special priors on the MDP's parameters, whereas the present work has no such requirements and can be applied in a more general context of RL methods.

The remainder of this paper is organised as follows. We first give an overview over Reinforcement Learning (sec. II) and uncertainty (sec. III). The main section IV discusses, how to bring these concepts together. We explain the application of uncertainty propagation to the Bellman iteration for policy evaluation and policy iteration for discrete MDPs and introduce the concept of certain-optimality. These approaches provide a general framework for different statistical paradigms, we particularly describe, how to apply

frequentist and Bayesian statistics. We further argue, that certain-optimal policies are stochastic in general. Since the approaches are applicable to function approximation as well, in sec. V the derivation to Least-Squares Policy-Iteration [17] is exemplarily elaborated. Finally, in sec. VI, we focus on artificial and industrial benchmarks and demonstrate different application domains.

II. Reinforcement Learning

In Reinforcement Learning the main objective is to achieve a policy, that optimally moves an agent within a Markov decision process, which is given by state and action spaces S and A as well as the dynamics, defined by a transition probability distribution $P_T : S \times A \times S \to [0, 1]$ depending on the the current state, the chosen action, and the successor state. The agent collects rewards while transiting, whose expected discounted future sum

$$V^\pi(s) \quad = \quad \mathrm{E}^\pi_{\mathbf{s}}\left(\sum_{i=0}^{\infty}\gamma^i R\left(s^{(i)}, \pi\left(s^{(i)}\right), s^{(i+1)}\right)\right),$$

the value function, has to be maximised over the policy space $\Pi \in (S \to A)$ for all possible states s, where $0 < \gamma < 1$ is the discount factor, s' the successor state of s, $\pi \in \Pi$ the used policy, and $\mathbf{s} = \left\{s', s'', \ldots, s^{(i)}, \ldots\right\}$. As an intermediate step one constructs a so-called Q-function $Q^\pi(s, a)$ depending on the current state and the chosen action. The optimal $Q^* = Q^{\pi^*}$ is determined by a solution of the Bellman optimality equation

$$\begin{aligned} Q^*(s,a) &= \mathrm{E}_{s'}(R(s,a,s') + \gamma V^*(s')) \\ &= \mathrm{E}_{s'}\Big(R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\Big). \end{aligned}$$

Therefore the best policy is $\pi^*(s) = \arg\max_a Q^*(s, a)$. We define the Bellman operator T as $(TQ)(s, a) = \mathrm{E}_{s'}(R(s, a, s') + \gamma \max_{a'} Q(s', a'))$ for any Q. The fixed point of $Q = \mathbf{Solve}(TQ)$, i.e. the Bellman operator followed by its projection on the Q-function's hypothesis space is the approached solution [1], [17], [4].

III. Uncertainty

Statistical uncertainty is a crucial issue in many application fields of statistics including the machine learning domain. It is well accepted that any measurement in nature and any conclusion from measurements are affected by an uncertainty. The International Organization for Standardization (ISO) defines uncertainty [18] to be "a parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand."

In the present work, we focus on the determination, quantisation, and minimisation of uncertainty of the measurements' conclusions in the context of Reinforcement Learning, i.e. the uncertainties of Q-functions and policies. The reason for uncertainty in RL is the ignorance about the true environment, i.e. the true MDP. The more observations are collected, the more certain the observer is about the MDP. And the larger the stochasticity, the more uncertainty remains about the MDP for a given amount of observations. And indeed, if the MDP is completely deterministic on the one hand, then everything is known about a state-action-pair, if it is observed once. There is no uncertainty left. If the system is, on the other hand, highly stochastic, then the risk of getting a low long-term return in expectation is large.

Note that the mentioned uncertainty is therefore qualitatively different from the MDP's stochasticity leading to the risk of obtaining a low long-term return in the single run. The main difference is, that the latter considers the inherent stochasticity of the MDP, whereas uncertainty considers the stochasticity of choosing an MDP from a set of MDPs.

The uncertainty of the measurements, i.e. the transitions and rewards, are propagated to the conclusions, e.g. the Q-function, by uncertainty propagation (UP), which is a common concept in statistics [2]. We determine the uncertainty of values $f(\mathrm{x})$ with $f : \mathbb{R}^m \to \mathbb{R}^n$ given the uncertainty of their arguments x as $\mathrm{Cov}(f) = \mathrm{Cov}(f, f) = D\mathrm{Cov}(\mathrm{x})D^T$, where $D_{i,j} = \frac{\partial f_i}{\partial \mathrm{x}_j}$ is the Jacobian matrix of f w.r.t. x and $\mathrm{Cov}(\mathrm{x}) = \mathrm{Cov}(\mathrm{x}, \mathrm{x})$ the covariance matrix of the arguments x holding the uncertainty of x, f then provides the (symmetric and positive definite) uncertainty $\mathrm{Cov}(f)$.

In the following, we usually work on multi-dimensional objects and on their covariance matrices. Therefore, those objects have to be appropriately vectorised. This can be done by any enumeration and is only of technical importance.

IV. Bellman Iteration and Uncertainty Propagation

Our concept of incorporating uncertainty into RL consists in applying UP to the Bellman iteration

$$\begin{aligned} Q^m(s_i,a_j) &= (TQ^{m-1})(s_i,a_j) \\ &= \sum_{k=1}^{|S|} P(s_k|s_i,a_j)\big(R(s_i,a_j,s_k) \\ &\quad +\gamma V^{m-1}(s_k)\big), \end{aligned}$$

here for discretised MDPs. For policy evaluation we have $V^m(s) = Q^m(s, \pi(s))$, with π the used policy and for policy iteration $V^m(s) = \max_{a \in A} Q^m(s, a)$ (sec. II). Thereby we assume a finite number of states $s_i, i \in \{1, \ldots, |S|\}$ and actions $a_i, i \in \{1, \ldots, |A|\}$. The Bellman iteration converges, with $m \to \infty$, to the (optimal) Q-function, which is appropriate to the estimators P and R. In the general stochastic case, which will be important later, we set $V^m(s) = \sum_{i=1}^{|A|} \pi(s, a_i)Q^m(s, a_i)$ with $\pi(s, a)$ the probability of choosing a in s. To obtain the uncertainty of the approached Q-function, the technique of UP is applied parallelly to the Bellman iteration. With given covariance matrices $\mathrm{Cov}(P)$, $\mathrm{Cov}(R)$, and $\mathrm{Cov}(P, R)$ for the transition probabilities and the rewards, we obtain the initial complete covariance matrix

$$\mathrm{Cov}(Q^0, P, R) \quad = \quad \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathrm{Cov}(P) & \mathrm{Cov}(P,R) \\ 0 & \mathrm{Cov}(P,R)^T & \mathrm{Cov}(R) \end{pmatrix}$$

and the complete covariance matrix after the mth Bellman iteration

$$\mathrm{Cov}(Q^m, P, R) \quad = \quad D^{m-1}\mathrm{Cov}(Q^{m-1}, P, R)(D^{m-1})^T$$

with the Jacobian matrix

$$\begin{aligned} D^m &= \begin{pmatrix} D^m_{Q,Q} & D^m_{Q,P} & D^m_{Q,R} \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix} \\ (D^m_{Q,Q})_{(i,j),(k,l)} &= \gamma\pi(s_k, a_l)P(s_k|s_i, a_j) \\ (D^m_{Q,P})_{(i,j),(l,n,k)} &= \delta_{il}\delta_{jn}\left(R(s_i, a_j, s_k) + \gamma V^m(s_k)\right) \\ (D^m_{Q,R})_{(i,j),(l,n,k)} &= \delta_{il}\delta_{jn}P(s_k|s_i, a_j). \end{aligned}$$

Note that the presented uncertainty propagation accords the expanded Bellman iteration

$$\begin{pmatrix} Q^m & P & R \end{pmatrix}^T \quad = \quad \begin{pmatrix} TQ^{m-1} & P & R \end{pmatrix}^T$$

to obtain the covariances between Q-function and P and R, respectively. All parameters of Q^m are linear in Q^m, altogether it is a bi-linear function. Therefore UP is indeed approximately applicable in this setting [2].

A. The Initial Covariances

The initial covariance matrix $\mathrm{Cov}((P, R))$ has to be designed by problem dependent prior belief. If e.g. all transitions from different state-action-pairs are assumed to be independent of each other and the rewards, all transitions can be modelled as multinomial distributions. In a Bayesian context one supposes a priorly known distribution [2], [19] over the parameter space $P(s_k|s_i, a_j)$ for given i and j. The Dirichlet distribution with density

$$\begin{aligned} &P(P(s_1|s_i, a_j), \ldots, P(s_{|S|}|s_i, a_j))_{\alpha_{1,i,j},\ldots,\alpha_{|S|,i,j}} \\ &= \frac{\Gamma(\alpha_{i,j})}{\prod_{k=1}^{|S|}\Gamma(\alpha_{k,i,j})}\prod_{k=1}^{|S|}P(s_k|s_i, a_j)^{\alpha_{k,i,j}-1} \end{aligned}$$

and $\alpha_{i,j} = \sum_{k=1}^{|S|}\alpha_{k,i,j}$ is a conjugate prior in this case with posterior parameters

$$\alpha^d_{k,i,j} \quad = \quad \alpha_{k,i,j} + n_{s_k|s_i,a_j}$$

in the light of the observations occurring $n_{s_k|s_i,a_j}$ times a transition from s_i to s_k by using action a_j. The initial covariance matrix for P then becomes

$$\begin{aligned} &(\mathrm{Cov}(P))_{(i,j,k),(l,m,n)} \\ &= \delta_{i,l}\delta_{j,m}\frac{\alpha^d_{k,i,j}(\delta_{k,n}\alpha^d_{i,j} - \alpha^d_{n,i,j})}{(\alpha^d_{i,j})^2(\alpha^d_{i,j}+1)}, \end{aligned}$$

assuming the posterior estimator $P(s_k|s_i, a_j) = \alpha^d_{k,i,j}/\alpha^d_{i,j}$. Similarly, the rewards might be distributed normally with the normal-gamma distribution as a conjugate prior.

As a simplification or by using the frequentist paradigm, it is also possible to use the relative frequency as the expected transition probabilities with their uncertainties

$$\begin{aligned} &(\mathrm{Cov}(P))_{(i,j,k),(l,m,n)} \\ &= \delta_{i,l}\delta_{j,m}\frac{P(s_k|s_i, a_j)(\delta_{k,n} - P(s_n|s_i, a_j))}{n_{s_i,a_j} - 1} \end{aligned}$$

with n_{s_i,a_j} observed transitions from the state-action-pair (s_i, a_j).

Similarly, the rewards expectations become their sample means and $\mathrm{Cov}(R)$ a diagonal matrix with entries

$$\mathrm{Cov}(R(s_i, a_j, s_k)) \quad = \quad \frac{\mathrm{Var}(R(s_i, a_j, s_k))}{n_{s_k|s_i,a_j} - 1}.$$

The frequentist view and the conjugate priors have the advantage of being computationally feasible, nevertheless, the method is not restricted to them, any meaningful covariance matrix $\mathrm{Cov}((P, R))$ is allowed. Particularly, applying covariances between the transitions starting from different state-action-pairs and between states and rewards is reasonable and interesting, if there is some measure of neighbourhood over the state-action-space. Crucial is finally, that the prior represents the user's belief.

Theorem 1: Suppose a finite MDP $M = (S, A, P, R)$ with discount factor $0 < \gamma < 1$ and C^0 an arbitrary initial symmetric and positive definite covariance matrix. Then the function

$$(Q^m, C^m) \quad = \quad \left(TQ^{m-1}, D^{m-1}C^{m-1}(D^{m-1})^T\right)$$

provides a unique fixed point (Q^*, C^*) almost surely, independent of the initial Q, for policy evaluation and policy iteration.

Proof: It has already been shown that $Q^m = TQ^{m-1}$ converges to a unique fixed point Q^* [1]. Since Q^m does not depend on C^k or D^k for any iteration $k < m$, Q^* persists. We obtain

$$C^m \quad = \quad \prod_{i=0}^{m-1}D^i C^0 \prod_{i=0}^{m-1}(D^i)^T$$

after m iterations. Due to convergence of Q^m, D^m converges to D^* as well, which leads to

$$C^* \quad = \quad \prod_{i=0}^{\infty}D^* C_{\mathrm{conv}} \prod_{i=0}^{\infty}(D^*)^T$$

with C_{conv} the covariance matrix after convergence of Q. By successive matrix multiplication we obtain

$$\begin{aligned} (D^*)^\infty &= \begin{pmatrix} (D^*)^\infty_{Q,Q} & \sum_{i=0}^{\infty}(D^*)^i_{Q,Q}(D^*)_{Q,P} \\ 0 & \mathrm{I} \\ 0 & 0 \end{pmatrix} \\ &\qquad \begin{matrix} \sum_{i=0}^{\infty}(D^*)^i_{Q,Q}(D^*)_{Q,R} \\ 0 \\ \mathrm{I} \end{matrix} \Bigg) \\ &= \begin{pmatrix} 0 & (\mathrm{I} - (D^*)_{Q,Q})^{-1}(D^*)_{Q,P} \\ 0 & \mathrm{I} \\ 0 & 0 \end{pmatrix} \\ &\qquad \begin{matrix} (\mathrm{I} - (D^*)_{Q,Q})^{-1}(D^*)_{Q,R} \\ 0 \\ \mathrm{I} \end{matrix} \Bigg) \end{aligned}$$

since all eigenvalues of $(D^*)_{Q,Q}$ are strictly smaller than 1 and $\mathrm{I} - (D^*)_{Q,Q}$ is invertible for all but finitely many

$(D^*)_{Q,Q}$. Therefore, almost surely, $(D^*)^\infty$ exists, which implies that C^* exists as well. We obtain finally

$$\begin{aligned} C^*_{Q,Q} &= (\mathrm{I} - (D^*)_{Q,Q})^{-1} \begin{pmatrix} (D^*)_{Q,P} & (D^*)_{Q,R} \end{pmatrix} \\ &\quad \begin{pmatrix} \mathrm{Cov}(P) & \mathrm{Cov}(P,R) \\ \mathrm{Cov}(P,R)^T & \mathrm{Cov}(R) \end{pmatrix} \\ &\quad \begin{pmatrix} (D^*)^T_{Q,P} \\ (D^*)^T_{Q,R} \end{pmatrix} \left((\mathrm{I} - (D^*)_{Q,Q})^{-1}\right)^T. \end{aligned}$$

The fixed point C^* depends on the initial covariance matrices $\mathrm{Cov}(P)$, $\mathrm{Cov}(R)$, and $\mathrm{Cov}(P,R)$ only and is therefore independent of the operations made before reaching the fixed point Q^*. □

Having identified the fixed point consisting of Q^* and its covariance $\mathrm{Cov}(Q^*)$ as part of C^*, the uncertainty of each individual state-action-pair is represented by the square root of the diagonal entries $\sigma Q^* = \sqrt{\mathrm{diag}(\mathrm{Cov}(Q^*))}$ since the diagonal comprises the Q-values' variances.

Finally, with probability $P(\xi)$ depending on the distribution class of Q, the function

$$Q^*_u(s,a) = (Q^* - \xi\sigma Q^*)(s,a)$$

provides the guaranteed performance expectation applying action a_j in state s_i strictly followed by the policy $\pi^*(s) = \arg\max_a Q^*(s,a)$. Suppose exemplarily Q to be distributed normally, then the choice $\xi = 2$ would lead to the guaranteed performance with $P(2) \approx 0.977$.

Note that this knowledge of uncertainty might help to guide exploration, i.e. to probe state-action-pairs with large uncertainty, but it does not help to improve the guaranteed performance in a principled manner. By applying $\pi(s) = \arg\max_a Q^*_u(s,a)$, the uncertainty would not be estimated correctly as the agent is only allowed once to decide for another action than the approached policy suggests.

B. Joint Iteration

To overcome this problem, we want to approach a so-called certain-optimal policy, which maximises the guaranteed performance. The idea is to obtain a policy π, which is optimal w.r.t. a specified confidence level, i.e. which maximises $Z(s,a)$ for all s and a such that

$$\forall s,a : P\left(\bar{Q}^\pi(s,a) > Z(s,a)\right) > P(\xi)$$

is fulfilled, where $\bar{Q}^\pi$ denotes the true performance function of π and $P(\xi)$ being a prespecified probability. We approach such a solution by approximating Z by Q^π_u and solving

$$\begin{aligned} \pi^\xi(s) &= \arg\max_\pi \max_a Q^\pi_u(s,a) \\ &= \arg\max_\pi \max_a \left(Q^\pi - \xi\sigma Q^\pi\right)(s,a) \end{aligned}$$

under the constraints, that $Q^{\pi^\xi} = Q^\xi$ is the valid Q-function for π^ξ, i.e.

$$\begin{aligned} \forall i,j : Q^\xi(s_i,a_j) &= \sum_{k=1}^{|S|} P(s_k|s_i,a_j)\Big(R(s_i,a_j,s_k) \\ &\quad +\gamma Q^\xi(s_k,\pi^\xi(s_k))\Big). \end{aligned}$$

Relating to the Bellman iteration, Q shall be a fixed point not w.r.t. the value function as the maximum over all Q-values, but the maximum over the Q-values minus its weighted uncertainty. Therefore, one has to choose

$$\pi^m(s) = \arg\max_{a'} (Q^m - \xi\sigma Q^m)(s,a')$$

after each iteration, together with an update of the uncertainties according to the modified policy π^m.

C. Certain-optimal Policies are Stochastic Policies

Policy evaluation can be made for deterministic and stochastic policies, whereas it has been proven that an optimal policy, within the framework of MDPs, is always deterministic [20]. For certain-optimal policies, however, the situation is different. Particularly, for $\xi > 0$ there is a bias on $\xi\sigma Q(s,\pi(s))$ being larger than $\xi\sigma Q(s,a), a \neq \pi(s)$, if π is the evaluated policy, since $R(s,\pi(s),s')$ depends stronger on $V(s') = Q(s',\pi(s'))$ than $R(s,a,s'), a \neq \pi(s)$. The value function implies the choice of action $\pi(s)$ for all further occurrences of state s. Therefore, the (deterministic) joint iteration is not necessarily guaranteed to converge. I.e., switching the policy π to π' with $Q(s,\pi'(s)) - \xi\sigma Q(s,\pi'(s)) > Q(s,\pi(s)) - \xi\sigma Q(s,\pi(s))$ could lead to a larger uncertainty of π' at s and hence to $Q'(s,\pi'(s)) - \xi\sigma Q'(s,\pi'(s)) < Q'(s,\pi(s)) - \xi\sigma Q'(s,\pi(s))$ for Q' at the next iteration. This causes an oscillation.

It is intuitively apparent that a certain-optimal policy should be stochastic in general if the gain in value must be balanced with the gain in certainty, i.e. with a decreasing risk of having estimated the wrong MDP. The risk to obtain a low expected return is hence reduced by diversification.

The value ξ decides about the cost of certainty. If $\xi > 0$ is large, certain-optimal policies tend to become more stochastic, one pays a price for the benefit of a guaranteed small performance, whereas a small $\xi \leq 0$ guarantees deterministic certain-optimal policies, however, uncertainty takes on the meaning of chance for a large performance. Therefore, we finally define a stochastic uncertainty incorporating Bellman iteration as

$$\begin{pmatrix} Q^m \\ C^m \\ \pi^m \end{pmatrix} = \begin{pmatrix} TQ^{m-1} \\ D_{m-1}C^{m-1}D^T_{m-1} \\ \Lambda(\pi^{m-1}, TQ^{m-1}, m) \end{pmatrix}$$

with

$$\begin{aligned} &\Lambda(\pi,Q,t)(s,a) \\ &= \begin{cases} \min\left(\pi(s,a) + \frac{1}{t}, 1\right) & : \quad a = a_Q(s) \\ \frac{\max\left(1-\pi(s,a_Q(s))-\frac{1}{t},0\right)}{1-\pi(s,a_Q(s))}\pi(s,a) & : \quad \text{otherwise} \end{cases} \end{aligned}$$

and $a_Q(s) = \arg\max_a (Q - \xi\sigma Q)(s,a)$. The harmonically decreasing change rate of the stochastic policies guarantees reachability of all policies on the one hand and convergence on the other hand. Alg. 1 summarises the joint iteration.

Note that the time complexity per iteration is of higher order than the standard Bellman iteration's one, which needs $O\left(|S|^2|A|\right)$ time. The bottleneck is the covariance update

Algorithm 1 Uncertainty Incorporating Joint Iteration for Discrete MDPs

Require: given estimators P and R for a discrete MDP, initial covariance matrices $\mathrm{Cov}(P)$, $\mathrm{Cov}(R)$, and $\mathrm{Cov}(P,R)$ as well as a scalar ξ

Ensure: calculates a certain-optimal Q-function Q and policy π under the assumption of the observations and the posteriors given by $\mathrm{Cov}(P)$, $\mathrm{Cov}(R)$, and $\mathrm{Cov}(P,R)$

set $C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mathrm{Cov}(P) & \mathrm{Cov}(P,R) \\ 0 & \mathrm{Cov}(P,R)^T & \mathrm{Cov}(R) \end{pmatrix}$

set $\forall i,j : Q(s_i,a_j) = 0$, $\forall i,j : \pi(s_i,a_j) = \frac{1}{|A|}$, $t = 0$

while the desired precision is not reached **do**

set $t = t+1$

set $\forall i,j : (\sigma Q)(s_i,a_j) = \sqrt{C_{i|A|+j,i|A|+j}}$

find $\forall i : a_{i,\max} = \arg\max_{a_j} (Q - \xi\sigma Q)(s_i,a_j)$

set $\forall i : d_{i,\mathrm{diff}} = \min\left(\frac{1}{t}, 1 - \pi(s_i,a_{i,\max})\right)$

set $\forall i : \pi(s_i,a_{i,\max}) = \pi(s_i,a_{i,\max}) + d_{i,\mathrm{diff}}$

set $\forall i : \forall a_j \neq a_{i,\max}$:

$\pi(s_i,a_j) = \frac{1-\pi(s_i,a_{i,\max})}{1-\pi(s,a_{i,\max})+d_{i,\mathrm{diff}}}\pi(s_i,a_j)$

set $\forall i,j : Q'(s_i,a_j) = \sum_{k=1}^{|S|} P(s_k|s_i,a_j)$

$\left(R(s_i,a_j,s_k) + \gamma \sum_{l=1}^{|A|} \pi(s_k,a_l) Q(s_k,a_l)\right)$

set $Q = Q'$

set $D = \begin{pmatrix} D_{Q,Q} & D_{Q,P} & D_{Q,R} \\ 0 & \mathrm{I} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}$

set $C = DCD^T$

end while

return $Q - \xi\sigma Q$ and π

with a time complexity between $O\left((|S||A|)^2 \log(|S||A|)\right)$ and $O\left((|S||A|)^{2.376}\right)$ [21] since each entry of Q depends only on $|S|$ entries of P and R. The overall complexity is hence bounded by these magnitudes.

The function $Q_u^\xi(s,a) = \left(Q^\xi - \xi\sigma Q^\xi\right)(s,a)$ with $\left(Q^\xi, C^\xi, \pi^\xi\right)$ as the fixed point of the (stochastic) joint iteration for given ξ provides, with probability $P(\xi)$ depending on the distribution class of Q, the guaranteed performance applying action a in state s strictly followed by the stochastic policy π^ξ. First and foremost, π^ξ maximises the guaranteed performance and is therefore called a certain-optimal policy.

V. Function Approximation

In order to demonstrate the principle functionality of the approach to any kind of function approximation, we apply the concept exemplarily on a variant of Least-Squares Policy-Iteration (LSPI) [17] for a finite number of actions. Here, the Q-function is linear and built upon features Φ as $Q_{\mathbf{w}}(s,a) = \Phi(s,a)^T\mathbf{w}$. The Bellman iteration then consists basically in iteratively solving the linear equation system

$$\begin{aligned} \left(\Phi^T\Phi + \lambda\mathrm{I}\right)\mathbf{w}^m &= \Phi^T\left(R + \gamma(V')^{m-1}\right) \\ &= \Phi^T\left(R + \gamma(\Phi')^{m-1}\mathbf{w}^{m-1}\right) \end{aligned}$$

for $\mathbf{w}^m$ with $\lambda \geq 0$ being a regularisation parameter, Φ the feature matrix of all observed state-action-pairs, and $(\Phi')^{m-1} = \sum_{a=1}^{|A|}\left(\pi^{m-1}(\cdot,a)\mathbf{1}^T\right) \bullet \Phi'_a$ with Φ'_a being the feature matrix of all successor states together with the action a, $\bullet$ the componentwise multiplication, $\mathbf{1}$ an appropriately sized vector consisting of ones, Q' and V' the Q-function and value function at the successor states.

The meaning and the information given by P and R in the discrete case are now provided by the observations, i.e. the successor states' features Φ' and the rewards $R = (r_1,\ldots,r_n)^T$ themselves, where n is the number of observations. Similarly we obtain the expanded Bellman iteration

$$\begin{pmatrix} \mathbf{w}^m \\ \begin{pmatrix} (Q'(\cdot,1))^{m-1} \\ \vdots \\ (Q'(\cdot,|A|))^{m-1} \end{pmatrix} \\ \Phi' \\ R \end{pmatrix} = \begin{pmatrix} \mathbf{v}^{m-1} \\ \begin{pmatrix} Z_1^{m-1} \\ \vdots \\ Z_{|A|}^{m-1} \end{pmatrix} \\ \Phi' \\ R \end{pmatrix},$$

where $Z_a^{m-1} = (\Phi'_a)^{m-1}\mathbf{v}^{m-1}$ and $\mathbf{v}^{m-1} = (\Phi^T\Phi + \lambda\mathrm{I})^{-1}\Phi^T(R + \gamma(V')^{m-1})$. The Jacobian matrix turns out to be

$$D^m = \begin{pmatrix} 0 & D^m_{\mathbf{w},Q'} & 0 & D^m_{\mathbf{w},R} \\ 0 & D^m_{Q',Q'} & D^m_{Q',\Phi'} & D^m_{Q',R} \\ 0 & 0 & \mathrm{I} & 0 \\ 0 & 0 & 0 & \mathrm{I} \end{pmatrix}$$

with entries

$$\begin{aligned} D^m_{\mathbf{w},(Q'(\cdot,a))} &= \gamma\left((\Phi^T\Phi + \lambda\mathrm{I})^{-1}\Phi^T\right) \bullet \left(\mathbf{1}\pi(\cdot,a)^T\right) \\ D^m_{\mathbf{w}^m,R} &= (\Phi^T\Phi + \lambda\mathrm{I})^{-1}\Phi^T \\ D^m_{Q'(\cdot,a),Q'(\cdot,b)} &= \gamma\left(\Phi'_a(\Phi^T\Phi + \lambda\mathrm{I})^{-1}\Phi^T\right) \\ &\quad \bullet \left(\mathbf{1}\pi(\cdot,b)^T\right) \\ D^m_{Q'(s_i,a),(\Phi'_b)_{j,\cdot}} &= \delta_{i,j}\delta_{a,b}\mathbf{w}^T \\ D^m_{Q'(\cdot,a),R} &= \Phi'_a(\Phi^T\Phi + \lambda\mathrm{I})^{-1}\Phi^T. \end{aligned}$$

We operate on the initial covariance matrix

$$\begin{aligned} &\mathrm{Cov}\left(\mathbf{w}^0, (Q')^0, \Phi', R\right) \\ &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \mathrm{Cov}(\Phi') & \mathrm{Cov}(\Phi',R) \\ 0 & 0 & \mathrm{Cov}(\Phi',R)^T & \mathrm{Cov}(R) \end{pmatrix}. \end{aligned}$$

Note that a convergence result as in th. 1 applies as well under the condition that $\Phi'_a(\Phi^T\Phi+\lambda\mathrm{I})^{-1}\Phi^T$ is a non-expansion for all actions a. The certain-optimal weight vector $\mathbf{w}_u$ can be determined by solving the linear program $(\Phi\mathbf{w}_u)^T\mathbf{1} \rightarrow \min$ under the constraints $\mathbf{w} - \xi(\sigma\mathbf{w}) \leq \mathbf{w}_u \leq \mathbf{w} + \xi(\sigma\mathbf{w})$. In a similar way, the method can also be applied to other approaches, which use function approximation together with the Bellman iteration.

Two problems remain. For the determination of the covariance matrix $\mathrm{Cov}((\Phi',R))$ a continuous generalisation of the transition probabilities' and rewards' distributions as well as their priors must be implemented. But in principle, the same as in the discrete case holds, it must represent the user's belief. Another issue is the handling of the policy's stochasticity, which could lead to stochastic weight vectors.

VI. Applications

The presented techniques offer at least four different types of applications, which are important in various practical domains.

A. Quality Assurance

With a positive ξ one aims at a guaranteed minimal performance of a given or the optimal policy. To optimise this minimal performance, we introduced the concept of and an approach to certain-optimality. The main practical motivation is to avoid delivering an inferior policy. To simply be aware of the quantification of uncertainty, helps to appreciate how well one can count on the result. If the guaranteed Q-value for a specified start state is insufficient, more observations must be provided in order to reduce the uncertainty.

If the exploration is expensive and the system critical, such that the performance probability has definitely to be fulfilled, it is reasonable to bring out the best from this concept. This can be achieved by a certain-optimal policy. One abandons optimality in order to perform as good as possible at the specified confidence level.

B. Exploration

Symmetrically, for negative ξ one uses the uncertainty in the opposite way by harnessing the chance for a high performance. This may be interesting to explore state-action-pairs, where $Q^{\xi}_{u}(s,a)$ is large, more intensively, since the estimator of the Q-value is already large but the true performance of the state-action-pair could be even better as the uncertainty is still large as well.

C. Competitions

Another application field are competitions, which is symmetrical to quality assurance by using negative ξ. The agent shall follow a policy, which gives him the chance to perform exceedingly well, and thus to win. In this case, certain-optimality comes again into play as the performance expectation is not the criterion, but the percentile performance.

For demonstration of the quality assurance and competition aspects, we applied the joint iteration on (fixed) data sets for two simple classes of MDPs (fig. 1). Subsequently we sampled over the space of allowed MDPs from their (fixed) prior distribution. As a result we achieve an accumulated posterior of the possible performances for each (stochastic) policy. Equipped with the correct priors, the approached policy indeed coincides approximately with the true certain-optimal policy.

Fig. 1 (left) concerns simple bandit problems with one state and two actions and fig. 1 (right) two-state MDPs with each two actions. The transition probabilities are assumed to be distributed multinomially for each start state, using the maximum entropy prior, i.e. the Beta distribution with $\alpha = \beta = 1$. For the rewards we assumed a normal distribution with fixed variance $\sigma_0 = 1$ and a normal prior for the mean with $\mu = 0$ and $\sigma = 1$. It can be seen that the certain-optimal policies are indeed stochastic for $\xi > 0$ and the performance estimation of the Q-function (dots and vertical lines) are close to the actual performance.

D. Increasing the Expected Performance

Incorporating uncertainty in RL can even improve the expected performance for concrete MDPs in many practical and industrial environments, where exploration is expensive and only allowed within a small range. The available amount of data is hence small and exploration takes place in an, in part extremely, unsymmetrical way. Data is particularly collected in areas, where the operation is already preferable. Many of the insufficiently explored so-called on-border-states are undesirable in expectation, but might, by chance, give a high reward in the singular case. If the border is sufficiently large, then it happens at least a few times, that such an outlier suggests a high expected reward. Note that in general the size of the border region will increase with the dimensionality of the problem. Carefully incorporating uncertainty avoids the agent to prefer those outliers in its final operation.

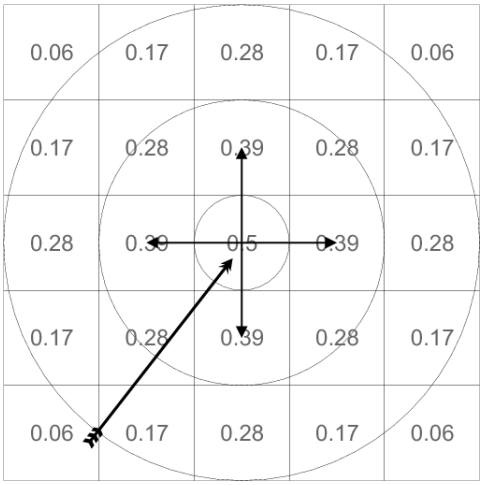

Fig. 2. Visualisation of the archery benchmark. The picture shows the target consisting of its 25 states, together with their hitting probabilities.

We applied the joint iteration on a simple artificial archery benchmark with the "border-phenomenon". The state space basically represents an archer's target (fig. 2). He possesses the possibility to move the arrowhead in all four directions and to shoot the arrow. The exploration has been performed randomly with short episodes. The dynamics were simulated with two different underlying MDPs. The arrowhead's moves are either stochastic (25 percent chance of choosing another action) or deterministic. The event of making a hit after shooting the arrow is stochastic in both settings. The highest probability for a hit is with the arrowhead in the target's middle. The border is explored quite rarely, such that a hit there misleadingly causes the respective estimator to estimate a high reward and thus the agent to finally shoot from this place.

In tbl. I the performance, averaged over 50 trials (two digits precision), for the frequentist setting (in the stochastic

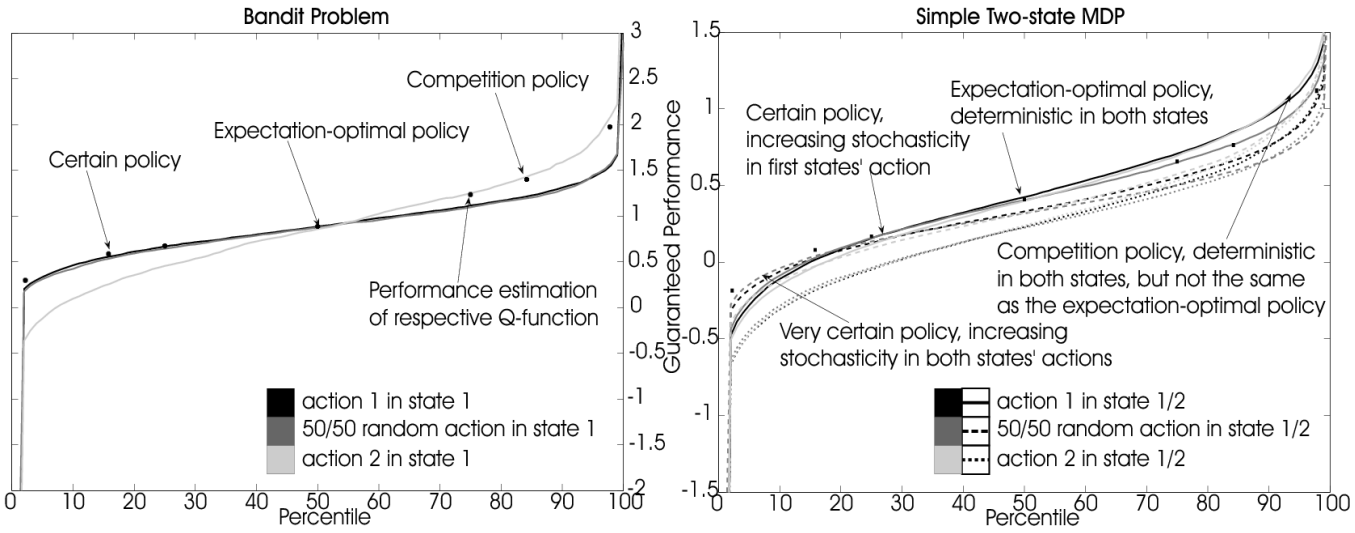

Fig. 1. Percentile performance for simple MDPs and joint iteration results. The different graphs show the minimal performances achieved by different (stochastic) policies. The grey scale value and the line style determine, which action to choose on the state/both states. The dots and vertical lines show the estimated Q-values for the certain-optimal policies at the specified percentile under the assumption, that those are distributed normally. Note that the plotted graphs are the inverse of the cumulative performance distributions. Since it is not possible to plot complete families of curves for each possible stochastic policy we chose each three representatives. This leads to 3 curves in the left figure and $3^2 = 9$ curves in the right one.

case) and the deterministic prior (in the deterministic case) for the transition probabilities are listed. Both priors can be implemented as Dirichlet distributions.

The table shows, that the performance indeed increases with ξ until a maximum and then decreases rapidly. The position of the maximum apparently increases with the number of observations. This can be explained by the decreasing uncertainty. The performance of the theoretical optimal policy is 0.31 for the stochastic archery benchmark and 0.5 for the deterministic one. They are achieved in average by the certain-optimal policy based on 2500 observations with $1 \leq \xi \leq 2$ in the stochastic case and for $3 \leq \xi \leq 4$ in the deterministic case.

E. Industrial Applications

We further applied the uncertainty propagation together with the joint iteration on an application to gas turbine control [22] with a continuous state and a finite action space, where it can be assumed, that the "border-phenomenon" appears as well. We discretised the internal state space with three different precisions (coarse ($4^4 = 256$ states), medium ($5^4 = 625$ states), fine ($6^4 = 1296$ states)), where the high-dimensional state space has already been reduced to a four-dimensional approximate Markovian state space. A detailed description of the problem and the construction of a minimal Markovian state space can be found in [22]. Note that the Bellman iteration and the uncertainty propagation is computationally feasible even with 6^4 states, since P and $\mathrm{Cov}((P, R))$ are sparse.

We summarise the averaged performances (50 trials with short random episodes starting from different operating points, lead to three digits precision) in tbl. I on the same uninformed priors as used in sec. VI-D. Those represent the frequentist view and the maximum entropy assumption, which leads to a uniform distribution, respectively. The rewards were estimated with an uninformed normal-gamma distribution as conjugate prior with $\sigma = \infty$ and $\alpha = \beta = 0$.

In contrary to the archery benchmark, we left the number of observations constant and changed the discretisation. The finer the discretisation, the larger is the uncertainty. Therefore the position of the maximum tends to increase with decreasing number of states. The performance is largest using the coarse discretisation. Indeed, averaged over all discretisations, the results for the frequentist prior tend to be better than for the maximum entropy prior. The overall best performance can be achieved with the coarse discretisation and the frequentist prior with $\xi = 5$, but using the maximum entropy prior leads to comparable results even with $\xi = 3$.

The theoretical optimum is not known, but for comparison we show the results of the Recurrent Q-Learning (RQL), Prioritised Sweeping (RPS), Fuzzy RL (RFuzzy), Neural Rewards Regression (RNRR), Policy Gradient NRR (RPGNRR), and Control Neural Network (RCNN), described in [22], [23], [24], respectively. The highest observed performance is 0.861 using 10^5 observations, which has almost been achieved by the best certain-optimal policy using 10^4 observations.

VII. Conclusion

A new approach to uncertainty incorporation in RL is presented. We applied the technique of UP to achieve certain-optimality and implemented it exemplarily on discretised MDPs and LSPI. Current and future work considers the application of our approach to other optimality criteria and function approximators as Neural Networks and Support Vector Machines. Also the application to further industrial environments is aspired. Another important issue is the utilisation of the information contained in the secondary diagonals of the covariance matrix. They are unused so far for both the decision, which action to select in the next

TABLE I

AVERAGE REWARD FOR THE ARCHERY AND GAS TURBINE BENCHMARK.

Setting	**Model**	**Discr.**	**# Obs.**	$\xi = \mathbf{0}$	$\xi = \frac{\mathbf{1}}{\mathbf{2}}$	$\xi = \mathbf{1}$	$\xi = \mathbf{2}$	$\xi = \mathbf{3}$	$\xi = \mathbf{4}$	$\xi = \mathbf{5}$
Archery	Frequentist		100	0.14	**0.16**	0.13	0.05	0.05	0.04	0.04
(Stochastic)	Dirichlet Prior		500	0.17	0.20	**0.25**	0.22	0.10	0.05	0.04
	$\forall i : \alpha_i = 0$		1000	0.21	0.26	**0.29**	0.27	0.22	0.11	0.07
			2500	0.27	0.29	**0.31**	**0.31**	0.30	0.28	0.24
Archery	Deterministic		100	0.35	**0.38**	0.23	0.17	0.12	0.11	0.09
(Deterministic)	Dirichlet Prior		500	0.32	0.38	0.39	**0.41**	0.27	0.18	0.11
	$\forall i : \alpha_i = 0$		1000	0.35	0.41	0.44	**0.45**	0.44	0.30	0.14
			2500	0.44	0.46	0.48	0.49	**0.50**	**0.50**	0.48
Turbine	Frequentist	coarse	10^4	0.736	0.758	0.770	0.815	0.837	0.848	**0.855**
	Dirichlet Prior	medium	10^4	0.751	0.769	0.784	0.816	**0.833**	0.830	0.815
	$\forall i : \alpha_i = 0$	fine	10^4	0.767	0.785	0.800	0.826	0.837	**0.840**	0.839
Turbine	Maximum Entropy	coarse	10^4	0.720	0.767	0.814	0.848	0.851	**0.854**	**0.854**
	Dirichlet Prior	medium	10^4	0.713	0.731	0.749	0.777	**0.787**	0.780	0.771
	$\forall i : \alpha_i = 1$	fine	10^4	0.735	0.773	0.789	**0.800**	**0.800**	0.786	0.779
	For Comparison			RefCon	RQL	RPS	RFuzzy	RNRR	RPGNRR	RCNN
Turbine		coarse	10^5		0.680	0.657	0.662			
		medium	10^5	0.53	0.687	0.745	0.657	0.851	0.861	0.859
		fine	10^5		0.717	0.729	0.668			

iteration step of the Bellman iteration and the operation of the final certain-optimal policy. One possibility is to switch from the local uncertainty measures to a global one. The guaranteed minimal performance with given probability for all or for at least one state, respectively, depends strongly on the covariances between the different states.

Definitely, as several laboratory conditions, such as the possibility of guided exploration or the access on a sufficiently large number of observations, are typically not fulfilled in practice, we conclude that the knowledge of uncertainty and its intelligent utilisation in Reinforcement Learning is vitally important to handle control problems of industrial scale.

REFERENCES

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 1998.

[2] Giulio D'Agostini. *Bayesian Reasoning in Data Analysis: A Critical Introduction*. World Scientific Publishing, 2003.

[3] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[4] Rémi Munos. Error bounds for approximate policy iteration. In *Proc. of the International Conference on Machine Learning*, pages 560–567, 2003.

[5] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. Approximate planning in large pomdps via reusable trajectories. In *Advances in Neural Information Processing Systems 12*, 2000.

[6] Leonid Peshkin and Sayan Mukherjee. Bounds on sample size for policy evaluation in Markov environments. In *14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory*, volume 2111, pages 616–629. Springer, Berlin, July 2001.

[7] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. In *Proc. of the Conference on Learning Theory*, pages 574–588, 2006.

[8] Yaakov Engel, Shie Mannor, and Ron Meir. Bayes meets Bellman: The gaussian process approach to temporal difference learning. In *Proc. of the International Conference on Machine Learning*, pages 154–161, 2003.

[9] Yaakov Engel, Shie Mannor, and Ron Meir. Reinforcement learning with gaussian processes. In *Proc. of the 22nd International Conference on Machine learning*, pages 201–208, 2005.

[10] Carl Edward Rasmussen and Malte Kuss. Gaussian processes in reinforcement learning. In *Advances in Neural Information Processing Systems 16*, pages 751–759, 2003.

[11] Mohammad Ghavamzadeh and Yaakov Engel. Bayesian policy gradient algorithms. In *Advances in Neural Information Processing Systems 19*, pages 457–464, 2006.

[12] Mohammad Ghavamzadeh and Yaakov Engel. Bayesian actor-critic algorithms. In *Proc. of the 24th International Conference on Machine learning*, pages 297–304, 2007.

[13] Richard Dearden, Nir Friedman, and Stuart J. Russell. Bayesian Q-learning. In *AAAI/IAAI*, pages 761–768, 1998.

[14] Richard Dearden, Nir Friedman, and David Andre. Model based bayesian exploration. In *Proc. Fifteenth Conf. on Uncertainty in Artificial Intelligence*, pages 150–159, 1999.

[15] Pascal Poupart, Nikos Vlassis, Jesse Hoey, and Kevin Regan. An analytic solution to discrete bayesian reinforcement learning. In *Proc. of the International Conference on Machine Learning*, pages 697 – 704, 2006.

[16] Erick Delage and Shie Mannor. Percentile optimization in uncertain markov decision processes with application to efficient exploration. In *Proc. of the International Conference on Machine Learning*, pages 225 – 232, 2007.

[17] Michael G. Lagoudakis and Ronald Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, pages 1107–1149, 2003.

[18] *Guide to the Expression of Uncertainty in Measurement*. International Organization for Standardization, 1993.

[19] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[20] Martin L. Puterman. *Markov Decision Processes*. John Wiley & Sons, New York, 1994.

[21] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9:251–280, 1990.

[22] Anton Maximilian Schaefer, Daniel Schneegass, Volkmar Sterzing, and Steffen Udluft. A neural reinforcement learning approach to gas turbine control. In *Proc. of the International Joint Conference on Neural Networks*, 2007.

[23] Martin Appl and Wilfried Brauer. Fuzzy model-based reinforcement learning. In *Advances in Computational Intelligence and Learning*, pages 211–223, 2002.

[24] Daniel Schneegass, Steffen Udluft, and Thomas Martinetz. Improving optimality of neural rewards regression for data-efficient batch near-optimal policy identification. In *Proc. of the International Conference on Artificial Neural Networks*, pages 109–118, 2007.