Batch Reinforcement Learning

Sascha Lange, Thomas Gabel, and Martin Riedmiller

Abstract Batch reinforcement learning is a subfield of dynamic programming-based reinforcement learning. Originally defined as the task of learning the best possible policy from a fixed set of a priori-known transition samples, the (batch) algorithms developed in this field can be easily adapted to the classical online case, where the agent interacts with the environment while learning. Due to the efficient use of collected data and the stability of the learning process, this research area has attracted a lot of attention recently. In this chapter, we introduce the basic principles and the theory behind batch reinforcement learning, describe the most important algorithms, exemplarily discuss ongoing research within this field, and briefly survey real-world applications of batch reinforcement learning.

1 Introduction

Batch reinforcement learning is a subfield of dynamic programming (DP) based reinforcement learning (RL) that has vastly grown in importance during the last years. Historically, the term 'batch RL' is used to describe a reinforcement learning setting, where the complete amount of learning experience—usually a set of transitions sampled from the system—is fixed and given a priori (Ernst et al, 2005a). The task of the learning system then is to derive a solution—usually an optimal policy—out of this given batch of samples.

In the following, we will relax this assumption of an a priori fixed set of training experience. The crucial benefit of batch algorithms lies in the way they handle a batch of transitions and get the best out of it, rather than in the fact that this set is fixed. From this perspective, batch RL algorithms are characterized by two basic constituents: all observed transitions are stored and updates occur synchronously on

Sascha Lange, Thomas Gabel, Martin Riedmiller
Albert-Ludwigs-Universtität Freiburg, Faculty of Engineering, Georges-Köhler-Allee 079, D-79110 Freiburg, Germany, e-mail: [slange,tgabel,riedmiller]@informatik.uni-freiburg.de

the whole batch of transitions ('fitting'). In particular, this allows for the definition of 'growing batch' methods, that are allowed to extend the set of sample experience in order to incrementally improve their solution. From the interaction perspective, the growing batch approach minimizes the difference between batch methods and pure online learning methods.

The benefits that come with the batch idea—namely, stability and data-efficiency of the learning process—account for the large interest in batch algorithms. Whereas basic algorithms like Q-learning usually need many interactions until convergence to good policies, thus often rendering a direct application to real applications impossible, methods including ideas from batch reinforcement learning usually converge in a fraction of the time. A number of successful examples of applying ideas originating from batch RL to learning in the interaction with real-world systems have recently been published (see sections 6.2 and 6.5).

In this chapter, we will first define the batch reinforcement learning problem and its variants, which form the problem space treated by batch RL methods. We will then give a brief historical recap of the development of the central ideas that, in retrospect, built the foundation of all modern batch RL algorithms. On the basis of the problem definition and the introduced ideas, we will present the most important algorithms in batch RL. We will discuss their theoretical properties as well as some variations that have a high relevance for practical applications. This includes a treatment of Neural Fitted Q Iteration (NFQ) and some of its applications, as it has proven a powerful tool for learning on real systems. With the application of batch methods to both visual learning of control policies and solving distributed scheduling problems, we will briefly discuss on-going research.

2 The Batch Reinforcement Learning Problem

Batch reinforcement learning was historically defined as the class of algorithms developed for solving a particular learning problem—namely, the batch reinforcement learning problem.

2.1 The Batch Learning Problem

As in the general reinforcement learning problem defined by Sutton and Barto (1998), the task in the batch learning problem is to find a policy that maximizes the sum of expected rewards in the familiar agent-environment loop. However, differing from the general case, in the batch learning problem the agent itself is not allowed to interact with the system during learning. Instead of observing a state s, trying an action a and adapting its policy according to the subsequent following state

s' and reward r, the learner only receives a set $\mathcal{F} = \{(s_t, a_t, r_{t+1}, s_{t+1}) | t = 1, \ldots, p\}$ of p transitions (s, a, r, s') sampled from the environment[1].

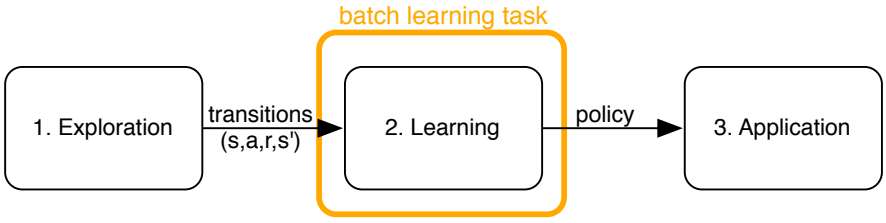

Fig. 1 The three distinct phases of the batch reinforcement learning process: 1: Collecting transitions with an arbitrary sampling strategy. 2: Application of (batch) reinforcement learning algorithms in order to learn the best possible policy from the set of transitions. 3: Application of the learned policy. Exploration is not part of the batch learning task. During the application phase, that isn't part of the learning task either, policies stay fixed and are not improved further.

In the most general case of this batch reinforcement learning problem the learner cannot make any assumptions on the sampling procedure of the transitions. They may be sampled by an arbitrary—even purely random—policy; they are not necessarily sampled uniformly from the sate-action space $S \times A$; they need not even be sampled along connected trajectories. Using only this information, the learner has to come up with a policy that will then be used by the agent to interact with the environment. During this application phase the policy is fixed and not further improved as new observations come in. Since the learner itself is not allowed to interact with the environment, and the given set of transitions is usually finite, the learner cannot be expected to always come up with an optimal policy. The objective has therefore been changed from learning an optimal policy—as in the general reinforcement learning case—to deriving the best possible policy from the given data.

The distinct separation of the whole procedure into three phases—exploring the environment and collecting state transitions and rewards, learning a policy, and application of the learned policy—their sequential nature, and the data passed at the interfaces is further clarified in figure 1. Obviously, treatment of the exploration–exploitation dilemma is not subject to algorithms solving such a pure batch learning problem, as the exploration is not part of the learning task at all.

[1] The methods presented in this chapter all assume a markovian state representation and the transitions $\mathcal{F}$ to be sampled from a discrete-time Markov decision process (MDP, see chapter **??**). For a treatment of only partially observable decision processes see chapter **??**.

2.2 The Growing Batch Learning Problem

Although this batch reinforcement learning problem historically has been the start of the development of batch reinforcement learning algorithms, modern batch RL algorithms are seldom used in this 'pure' batch learning problem. In practice, exploration has an important impact on the quality of the policies that can be learned. Obviously, the distribution of transitions in the provided batch must resemble the 'true' transition probabilities of the system in order to allow the derivation of good policies. The easiest way to achieve this is to sample the training examples from the system itself, by simply interacting with it. But when sampling from the real system, another aspect becomes important: the covering of the state space by the transitions used for learning. If 'important' regions—e.g. states close to the goal state—are not covered by any samples, then it is obviously not possible to learn a good policy from the data, since important information is missing. This is a real problem because in practice, a completely 'uninformed' policy—e.g. a purely random policy—is often not able to achieve an adequate covering of the state space—especially in the case of attractive starting states and hard to reach desirable states. It is often necessary to already have a rough idea of a good policy in order to be able to explore interesting regions that are not in the direct vicinity of the starting states.

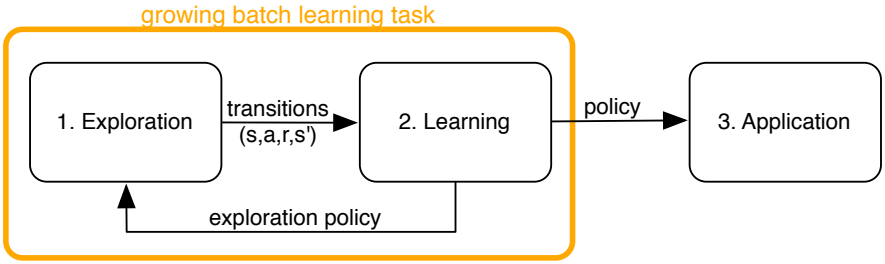

Fig. 2 The growing batch reinforcement learning process has the same three phases as the 'pure' batch learning process depicted in figure 1. But differing from the pure batch process, the growing batch learning process alternates for several times between the exploration and the learning phase, thus incrementally 'growing' the batch of stored transitions using intermediate policies.

This is the main reason why a third variant of the reinforcement learning problem became a popular practice, somehow positioned in between the pure online problem and the pure batch problem. Since the main idea of this third type of learning problem is to alternate between phases of exploration, where a set of training examples is grown by interacting with the system, and phases of learning, where the whole batch of observations is used (see fig. 2), we will refer to it as the 'growing batch' learning problem. In the literature, this growing batch approach can be found in several different guises; the number of alternations between episodes of exploration and episodes of learning can be in the whole range of being as close to the pure batch approach as using only two iterations (Riedmiller et al, 2008) to recal-

culating the policy after every few interactions—e.g. after finishing one episode in a shortest-path problem (Kalyanakrishnan and Stone, 2007; Lange and Riedmiller, 2010a). In practice, the growing batch approach is the modeling of choice when applying batch reinforcement learning algorithms to real systems. Since from the interaction perspective the growing batch approach is very similar to the 'pure' online approach—the agent improves its policy *while* interacting with the system—the interaction perspective, with the distinction between 'online' and 'offline', isn't that useful anymore for identifying batch RL algorithms. When talking about 'batch' RL algorithms now, it's more important to look at the algorithms and search for typical properties of the specific update rules.

3 Foundations of Batch RL Algorithms

Model-free online learning methods like Q-learning are appealing from a conceptual point of view and have been very successful when applied to problems with small, discrete state spaces. But when it comes to applying them to more realistic systems with larger and, possibly, continuous state spaces, these algorithms come up against limiting factors. In this respect, there can be identified three independent problems:

1. the 'exploration overhead', causing slow learning in practice
2. inefficiencies due to the stochastic approximation
3. stability issues when using function approximation

A common factor in modern batch RL algorithms is that these algorithms typically address all three issues and come up with specific solutions to each of them. In the following, we will discuss these problems in more detail and present the proposed solutions (in historical order) that now form the defining ideas behind modern batch reinforcement learning.

The Idea of 'Experience Replay' for Addressing the Exploration Overhead

In order to explain the problem referred to as 'exploration overhead', let us for a moment consider the common Q-update rule for model-free online learning with learning rate α and discount factor γ as given by

$$Q'(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a' \in A} Q(s',a') - Q(s,a) \right] \tag{1}$$

(see chapter **??**). In pure online Q-learning the agent alternates between learning and exploring with practically every single time step: in state s the agent selects and executes an action a, and then, when observing the subsequent state s' and reward r, it immediately updates the value function (and thus the corresponding greedy policy; see chapter **??**) according to (1), afterwards forgetting the 'experienced' state transition tuple (s,a,r,s'). It then returns to exploring with the updated policy

$Q'(s,a)$. Although this approach is guaranteed to converge in the limit, there is a severe performance problem with these 'local' updates along actually experienced transitions. When, for example, updating the Q-value of state-action pair (s_t, a_t) in time step t this may influence the values (s_{t-1}, a) for all $a \in A$ of a preceding state s_{t-1}. However, this change will not back-propagate immediately to all the involved preceding states, as the states preceding s_t are only updated the next time they are visited. And states preceding those preceding states s_{t-1} need yet another trial afterwards to be adapted to the new value of s_{t-1}. The performance problem with this is that these interactions are not actually needed to collect more information from the system in the first place, but mainly are needed to spread already available information through the whole state space in reverse order along trajectories. In practice, many interactions in Q-learning are of this 'spreading' type of interaction. This problem becomes even more pronounced when considering that those updates in model-free Q-learning are 'physical'—in the sense of needing real interaction with the system—and can seldomly be sampled in ordered sweeps across the whole state space or, at least, according to a uniform distribution.

In order to overcome this performance issue, Lin (1992) introduced the idea of 'experience replay'. Although the intention was to solve the 'exploration overhead' problem in online learning, in retrospect, we might identify it as basic—but nevertheless first—technique used for addressing the growing batch problem. The idea behind experience replay is to speed up convergence not only by using observed state transitions (the experience) once, but replaying them repeatedly to the agent as if they were new observations collected while interacting with the system. In fact, one can store a few to as many as all transitions observed up until that point and use them to update the Q-function for every single stored transition according to (1) after every interaction with the system. It is exaclty the information in the stored transitions that is missing in basic online Q-learning to be able to back-propagate information from updated states to preceding states without further interaction. The transitions collected when using experience replay resemble the connection between individual states and make more efficient use of this information, by spreading it along these connections, and, ideally, speeding up convergence.

The Idea of 'Fitting' to Address Stability Issues

In online RL it is common to use 'asynchronous' updates in the sense that the value function after each observation is immediately updated locally for that particular state, leaving all other states untouched. In the discrete case, this means updating one single Q-value for a state-action pair (s,a) in Q-learning—thereby immediately 'over-writing' the value of the starting state of the transition. Subsequent updates would use this updated value for their own update. This idea was also used with function approximation, first performing a DP update like, for example,

$$\bar{q}_{s,a} = r + \gamma \max_{a' \in A} f(s', a') \,, \tag{2}$$

recalculating the approximated value $\bar{q}_{s,a}$ of the present state action pair (s,a) by, e.g., adding immediate reward and approximated value of the subsequent state, and then immediately 'storing' the new value in the function approximator in the sense of moving the value of the approximation slightly towards the value of the new estimate $\bar{q}_{s,a}$ for the state-action pair (s,a):

$$f'(s,a) \leftarrow (1-\alpha)\, f(s,a) + \alpha\, \bar{q}_{s,a}\ . \tag{3}$$

Please note that (2) and (3) are just re-arranged forms of equation (1), using an arbitrary function approximation scheme for calculating an approximation f' of the 'true' Q-value function Q'.

Baird (1995), Gordon (1996) and others have shown examples where particular combinations of Q-learning and similar updates with function approximators behave instably or even lead to sure divergence. Stable behavior can be proven only for particular instances of combinations of approximation schemes and update rules, or under particular circumstances and assumptions on the system and reward structure (Schoknecht and Merke, 2003). In practice it required extensive experience on behalf of the engineer, in order to get a particular learning algorithm to work on a system. The observed stability issues are related to the interdependency of errors made in function approximation and deviations of the estimated value function from the optimal value function. Whereas the DP update (2) tries to gradually decrease the difference between $Q(s,a)$ and the optimal Q-function $Q^*(s,a)$, storing the updated value in the function approximator in step (3) might (re-)introduce an even larger error. Moreover, this approximation error influences all subsequent DP updates and may work against the contraction or even prevent it. The problem becomes even worse when global function approximators—like, for example, multi-layer perceptrons (Rumelhart et al, 1986; Werbos, 1974)—are used; improving a single Q-value of a state-action pair might impair all other approximations throughout the entire state space.

In this situation Gordon (1995a) came up with the compelling idea of slightly modifying the update scheme in order to separate the dynamic programming step from the function approximation step. The idea is to first apply a DP update to all members of a set of so-called 'supports' (points distributed throughout the state space) calculate new 'target values' for all supports (like in (2)), and then use supervised learning to train ('fit') a function approximator on all these new target values, thereby replacing the local updates of (3) (Gordon, 1995a). In this sense, the estimated Q-function is updated 'synchronously', with updates occuring at all supports at the same time. Although Gordon introduced this fitting idea within the setting of model-based value iteration, it became the foundation, and perhaps even the starting point, of all modern batch algorithms.

Replacing Inefficient Stochastic Approximation

Gordon discussed the possibility of transfering the 'fitting' idea to model-free—sample-based—approaches like, e.g., Q-learning, but did not find a solution and identified several convergence issues when estimating the values at the supports from samples in their surrounding. Ormoneit and Sen finally came up with the solution of how to adapt and apply his idea to the sample-based case. In their work, Ormoneit and Sen (2002) propose not to use arbitrarily selected supports in the state space to approximate value functions, but rather to use the sampled transitions directly to approximate the value function—either at the starting or at the ending states of the transitions—with the help of a kernel-based approximator. Their idea is to calculate an estimate of the value of each explored state-action pair under the actually observed transition (reward plus expected value of subsequent state), and then estimate the values of the subsequent states by averaging over the values of nearby transitions. Technically, the trick is to replace the exact DP-operator that was introduced by Gordon with a non-exact random operator (see section 4.1 for a formal definition). This random operator does not use the exact models in the DP-step. Instead, it only estimates the exact DP-operation by using a random sampling of transitions drawn from the unknown transition models. Lagoudakis and Parr (2001, 2003) came up with a similar idea independently.

As a side effect, calculating the costs of particular transitions and estimating the values of states (and actions) by averaging over the stored samples solved another performance problem connected to the stochastic approximation in regular Q-learning. Whereas in model-based value iteration the value of a state according to the current values of its possible subsequent states can be updated in a single step using the transition model, in Q-learning—due to the replacement of the model by stochastic approximation—such an update requires many visits to the state in question. Moreover, since we are using one learning rate for the entire state space, this learning rate cannot be adjusted in a way that is optimal for all states depending on the number of visits. In practice, this makes reaching the optimal convergence rate impossible. The algorithm of Ormoneit and Sen does not rely on stochastic approximation, but rather implicitly estimates the transition model by averaging over the observed transitions that—if collected in interaction with the system—actually form a random sampling from the true distributions.

4 Batch RL Algorithms

In their work, Ormoneit and Sen proposed a 'unified' kernel-based algorithm that could actually be seen as a general framework upon which several later algorithms are based. Their 'kernel-based approximate dynamic programming' (KADP) brought together the ideas of experience replay (storing and re-using experience), fitting (separation of DP-operator and approximation), and kernel-based self-approximation (sample-based).

4.1 Kernel-Based Approximate Dynamic Programming

Ormoneit's kernel-based approximate dynamic programming solves an approximated version of the 'exact' Bellman-equation

$$V = HV$$

that is expressed in

$$\hat{V} = \hat{H}\hat{V}\,.$$

It not only uses an approximation $\hat{V}$ of the 'true' state-value function V—as Gordon's fitted value iteration did—but also uses an approximate version $\hat{H}$ of the exact DP-operator H itself.

The KADP algorithm for solving this equation works as follows: starting from an arbitrary initial approximation $\hat{V}^0$ of the state-value function, each iteration i of the KADP-algorithm consists of solving the equation

$$\hat{V}^{i+1} = H_{max}\hat{H}^a_{dp}\hat{V}^i$$

for a given set

$$\mathscr{F} = \{(s_t, a_t, r_{t+1}, s_{t+1}) | t = 1, \ldots, p\}$$

of p transitions (s,a,r,s'). In this equation, the approximate DP-operator

$$\hat{H} = H_{max}\hat{H}^a_{dp}$$

has been split into an exact part H_{max} maximizing over actions and an approximate random operator $\hat{H}^a_{dp}$ approximating the 'true' (model-based) DP-step for individual actions from the observed transitions. The first half of this equation is calculated according to the sample-based DP update

$$\hat{Q}^{i+1}_a(\sigma) := \hat{H}^a_{dp}\hat{V}^i(\sigma) = \sum_{(s,a,r,s') \in \mathscr{F}_a} k(s,\sigma)\left[r + \gamma \hat{V}^i(s')\right]\,. \tag{4}$$

Using a weighting-kernel $k(\,\cdot\,,\sigma)$, this equation calculates a weighted average of the well-known Q-updates

$$r + \gamma \hat{V}^i(s') = r + \gamma \max_{a' \in A} \hat{Q}^i(s', a')\,,$$

(please note the similarity to equation (2)) along all transitions $(s,a,r,s') \in \mathscr{F}_a$, where $\mathscr{F}_a \subset \mathscr{F}$ is the subset of $\mathscr{F}$ that contains only the transitions $(s,a,r,s') \in \mathscr{F}$ that used the particular action a. The weighting kernel is chosen in such a way that more distant samples have a smaller influence on the resulting sum than closer (= more similar) samples.

The second half of the equation applies the maximizing-operator H_{max} to the approximated Q-functions $\hat{Q}^{i+1}_a$:

$$\hat{V}^{i+1}(s) = H_{max}\hat{Q}_a^{i+1}(s) = \max_{a\in A} \hat{Q}_a^{i+1}(s)\,. \tag{5}$$

Please note that this algorithm uses an individual approximation $\hat{Q}_a^{i+1} : S \mapsto R$ for each action $a \in A$ in order to approximate the Q-function $Q_a^{i+1} : S \times A \mapsto R$. Furthermore, a little bit counter-intuitively, in a practical implementation, this last equation is actually evaluated and stored for all ending states s' of all transitions $(s,a,r,s') \in \mathcal{F}$—not the starting states s. This decision to store the ending states is explained by noting that in the right-hand side of equation (4) we only query the present estimate of the value function at the ending states of the transitions—never its starting states (see Ormoneit and Sen (2002), section 4). This becomes clearer in figure 3.

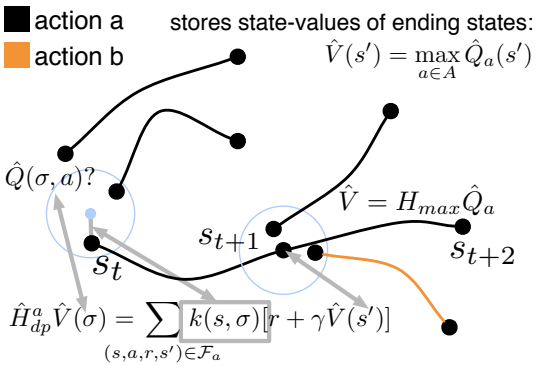

Fig. 3 Visualization of the kernel-based approximation in KADP. For computing the Q-value $\hat{Q}(\sigma,a) = \hat{H}^a_{dp}(\sigma)\hat{V}(\sigma)$ of (σ,a) at an arbitrary state $\sigma \in S$, KADP uses the starting states s of nearby transitions (s,a,r,s') only for calculating the weighting factors $k(s,\sigma)$, however, depends on the state values $\hat{V}(s')$ of ending states s' (in the depicted example $s = s_t$ and $s' = s_{t+1}$).

Iterating over equations (4) and (5), calculating a sequence of approximations $\hat{V}^{i+1}$ with $i = 0,1,2,\ldots$, the algorithm will ultimately converge to a unique solution $\hat{V}$ when explicitly storing the values $\hat{V}^i(s')$ of $\hat{V}^i$ at the end-points s' of the transitions (s,a,r,s') and using weighting-kernels that adhere to the 'averager' restriction from Gordon (1995a). For meeting this restriction the weights need to, A), add up to one

$$\sum_{(s,a,r,s')\in\mathcal{F}_a} k(s,\sigma) = 1 \quad \forall\sigma \in S$$

and, B), to be non-negative

$$k(s,\sigma) \geq 0 \quad \forall\sigma \in S \;\; \forall(s,a,r,s') \in \mathcal{F}_a\,.$$

The decision to store and iterate over state-values instead of state-action values results in the necessity of applying another DP-step in order to derive a greedy policy from the result of the KADP algorithm:

$$\begin{aligned}\pi^{i+1}(\sigma) &= \arg\max_{a\in A} \hat{H}^a_{dp}\hat{V}^i(\sigma) \\ &= \arg\max_{a\in A} \sum_{(s,a,r,s')\in\mathscr{F}_a} k(s,\sigma)\left[r+\gamma\hat{V}^i(s')\right] .\end{aligned}$$

This might appear to be a small problem from an efficiency point of view, as you have more complex computations as in Q-learning and need to remember all transitions in the application phase, but it is not a theoretical problem. Applying the DP-operator to the fixed point of $\hat{V} = \hat{H}\hat{V}$ does not change anything but results in the same unique fixed point.

4.2 Fitted Q Iteration

Perhaps the most popular algorithm in batch RL is Damien Ernst's 'Fitted Q Iteration' (FQI, Ernst et al (2005a)). It can be seen as the 'Q-Learning of batch RL', as it is actually a straight-forward transfer of the basic Q-learning update-rule to the batch case. Given a fixed set $\mathscr{F} = \{(s_t, a_t, r_{t+1}, s_{t+1}) | t = 1,\ldots,p\}$ of p transitions (s,a,r,s') and an initial Q-value $\bar{q}^0$ (Ernst et al (2005a) used $\bar{q}^0 = 0$) the algorithm starts by initializing an initial approximation $\hat{Q}^0$ of the Q-function Q^0 with $\hat{Q}^0(s,a) = \bar{q}^0$ for all $(s,a) \in S\times A$. It then iterates over the following two steps:

1. Start with an empty set P^{i+1} of patterns $(s,a;\bar{q}^{i+1}_{s,a})$. Then, for each transition $(s,a,r,s') \in \mathscr{F}$ calculate a new target Q-value $\bar{q}^{i+1}_{s,a}$ according to

$$\bar{q}^{i+1}_{s,a} = r + \gamma \max_{a'\in A} \hat{Q}^i(s',a') \tag{6}$$

   (similar to the update in equation (2)) and add a corresponding pattern $(s,a;\bar{q}^{i+1}_{s,a})$ to the pattern set; thus:

$$P^{i+1} \leftarrow P^{i+1} \cup \{(s,a;\bar{q}^{i+1}_{s,a})\} .$$

2. Use supervised learning to train a function approximator on the pattern set P^{i+1}. The resulting function $\hat{Q}^{i+1}$ then is an approximation of the Q-function Q^{i+1} after $i+1$ steps of dynamic programming.

Originally, Ernst proposed randomized trees for approximating the value function. After fixing their structure, these trees can also be represented as kernel-based averagers, thereby reducing step 2 to

$$\hat{Q}^{i+1}_a(\sigma) = \sum_{(s,a;\bar{q}^{i+1}_{s,a})\in P^{i+1}_a} k(s,\sigma)\bar{q}^{i+1}_{s,a} , \tag{7}$$

with the weights $k(\cdot,\sigma)$ determined by the structure of the tree. This variant of FQI constructs an individual approximation $\hat{Q}^{i+1}_a$ for each discrete action $a \in A$ which, together, form the approximation $\hat{Q}^{i+1}(s,a) = \hat{Q}^{i+1}_a(s)$ (Ernst et al, 2005a, section

3.4). Besides this variant of FQI, Ernst also proposed a variant with continuous actions. We may refer the interested reader to Ernst et al (2005a) for a detailed description of this.

From a theoretical stand-point, Fitted Q Iteration is nevertheless based on Ormoneit and Sen's theoretical framework. The similarity between Fitted Q Iteration and KADP becomes obvious when rearranging equations (6) and (7):

$$\hat{Q}^{i+1}(\sigma,a) = \hat{Q}_a^{i+1}(\sigma) = \sum_{(s,a;\bar{q})\in P_a^{i+1}} k(s,\sigma)\bar{q}_{s,a}^{i+1} \tag{8}$$

$$= \sum_{(s,a,r,s')\in \mathcal{F}_a} k(s,\sigma)\left[r+\gamma \max_{a'\in A} \hat{Q}_{a'}^{i}(s')\right] . \tag{9}$$

Equation (8) is the original averaging step from equation (7) in FQI for discrete actions. By inserting FQI's DP-step (6) immediately follows (9). This result (9) is practically identical to the update used in KADP, as can be seen by inserting (5) into (4):

$$\hat{Q}_a^{i+1}(\sigma) = \sum_{(s,a,r,s')\in \mathcal{F}_a} k(s,\sigma)\left[r+\gamma \hat{V}^{i}(s')\right]$$

$$= \sum_{(s,a,r,s')\in \mathcal{F}_a} k(s,\sigma)\left[r+\gamma \max_{a'\in A} \hat{Q}_{a'}^{i}(s')\right] .$$

Besides the optional treatment of continuous actions, another difference between FQI and KADP is in the splitting of operators and choice of explicitly represented values. Where KADP explicitly represents and uses state-values in the DP-step (4), FQI explicitly represents the Q-function and calculates state-values $\hat{V}^i(s) = max_{a\in A}\hat{Q}^i(s,a)$ on the fly by maximizing over actions in its DP-step (6). Although 'lazy-learning' with kernel-based averaging—as proposed as the standard in KADP—is also allowed in FQI, Ernst assumes the usage of a trained averager or other parametric function approximator for explicitly storing the Q-function. The 'structure' of this approximator is neither necessarily related to the starting points nor ending points of transitions. And, since FQI represents the Q-function explicitly, deriving a greedy policy in FQI is rather simple:

$$\pi^i(s) = \arg\max_{a\in A} \hat{Q}^i(s,a)$$

where $\hat{Q}^i$ is realized explicitly, either by a single function approximator (continuous actions) or by a set of function approximators $\hat{Q}_a^i$ for the actions $a \in A$.

These subtle differences in otherwise equivalent algorithms make FQI both, more intuitive and more similar to the online approach of Q-learning. This may account for this algorithm's greater popularity.

4.3 Least-Squares Policy Iteration

Least-squares policy iteration (LSPI, Lagoudakis and Parr (2003)) is another early example of a batch mode reinforcement learning algorithm. In contrast to the other algorithms reviewed in this section, LSPI explicitly embeds the task of solving control problems into the framework of policy iteration (Sutton and Barto, 1998), thus alternating between policy evaluation and policy improvement steps. However, LSPI never stores a policy explicitly. Instead, it works solely on the basis of a state-action value function Q from which a greedy policy is to be derived via $\pi(s) = \arg\max_{a \in A} Q(s,a)$. For the purpose of representing state-action value functions, LSPI employs a parametric linear approximation architecture with a fixed set of k pre-defined basis functions $\phi_i : S \times A$ and a weight vector $w = (w_1, \ldots, w_k)^T$. Therefore, any approximated state-action value function $\hat{Q}$ within the scope of LSPI takes the form

$$\hat{Q}(s,a,;w) = \sum_{j=1}^{k} \phi_j(s,a) w_j = \Phi w^T.$$

Its policy evaluation step employs a least-squares temporal difference learning algorithm for the state-action value function (LSQ, Lagoudakis and Parr (2001), later called LSTDQ, Lagoudakis and Parr (2003)). This algorithm takes as input the current policy π_m—as pointed out above, represented by a set of weights that determine a value function $\hat{Q}$ from which π_m is to be derived greedily—as well as a finite set $\mathscr{F}$ of transitions (s,a,r,s'). From these inputs, LSTDQ derives analytically the state-action value function $\hat{Q}^{\pi_m}$ for the given policy under the state distribution determined by the transition set. Clearly, the derived value function returned $\hat{Q}^{\pi_m}$ by LSTDQ is, again, fully described by a weight vector w^{π_m} given the above-mentioned linear architecture used.

Generally, the searched for value function is a fixed point of the $\hat{H}_\pi$ operator

$$(\hat{H}_\pi Q)(s,a) = r(s,\pi(s)) + \gamma \sum_{s' \in S} p(s,\pi(s),s') \max_{b \in A} Q(s,b), \tag{10}$$

i.e. $\hat{H}_{\pi_m} Q^\pi = Q^{\pi_m}$. Thus, a good approximation $\hat{Q}^{\pi_m}$ should comply to $\hat{H}_{\pi_m} \hat{Q}^{\pi_m} \approx \hat{Q}^{\pi_m} = \Phi (w^{\pi_m})^T$. Practically speaking, LSTDQ aims at finding a vector w^π such that the approximation of the result of applying the $\hat{H}_{\pi_m}$ operator to $\hat{Q}_{\pi_m}$ is as near as possible to the true result (in an L_2 norm minimizing manner). For this, LSTDQ employs an orthogonal projection and sets

$$\hat{Q}^{\pi_m} = (\Phi(\Phi^T \Phi)^{-1} \Phi^T) \hat{H} \hat{Q}^{\pi_m}. \tag{11}$$

With the compactly written version of equation (10), $\hat{H}_{\pi_m} Q^{\pi_m} = \mathscr{R} + \gamma \mathscr{P} \Pi_{\pi_m} Q^{\pi_m}$, equation (11) can be rearranged to

$$w^{\pi_m} = (\Phi^T (\Phi - \gamma \mathscr{P} \Pi_{\pi_m} \Phi))^{-1} \Phi^T \mathscr{R}$$

where Π_{π_m} is a stochastic matrix of size $|S| \times |S||A|$ describing policy π_m:

$$\Pi_{\pi_m}(s,(s',a')) = \pi_m(s',a').$$

Importantly, in this equation LSTDQ approximates the model of the system on the basis of the given sample set $\mathscr{F}$, i.e., P is a stochastic matrix of size $|S||A| \times |S|$ that contains transition probabilities, as observed within the transition set according to

$$P((s,a),s') = \sum_{(s,a,\cdot,s') \in \mathscr{F}} 1 / \sum_{(s,a,\cdot,\cdot) \in \mathscr{F}} 1 \approx Pr(s,a,s')$$

and $\mathscr{R}$ is a vector of size $|S||A|$ that summarizes the rewards contained in $\mathscr{F}$.

After having determined the state-action value function $\hat{Q}^{\pi_m}$ for the current policy π_m, a greedy (improved) policy π_{m+1} can be derived as usual by letting

$$\pi_{m+1}(s) = \arg\max_{a \in A} \hat{Q}^{\pi}_m(s,a) = \arg\max_{a \in A} \phi(s,a)(w^{\pi_m})^T.$$

Since LSPI never stores policies explicitly, but rather implicitly by the set of basis functions and corresponding weights—and thus, in fact, by a state-action value function—the policy improvement step of LSPI merely consists of overwriting the old weight vector w with the current weight vector found by a call to LSTDQ.

Valuable insight can be obtained by comparing a single iteration of the Fitted Q Iteration algorithm with a single iteration (one policy evaluation and improvement step) of LSPI. The main difference between LSPI and value function-centered FQI algorithms is that, in a single iteration, LSPI determines an approximation of the state-action value function Q^{π_m} for the current policy and batch of experience. LSPI can do this analytically—i.e., without iterating the $\hat{H}_{\pi_m}$ operator—because of the properties of its linear function approximation architecture. By contrast, FQI algorithms rely on a set of target values for the supervised fitting of a function approximator that are based on a single dynamic programming update step—i.e., on a single application of the $\hat{H}$ operator. Consequently, if we interpret Fitted Q Iteration algorithms from a policy iteration perspective, then this class of algorithms implements a batch variant of optimistic policy iteration, whereas LSPI realizes standard (non-optimistic) policy iteration.

4.4 Identifying Batch Algorithms

Whereas the algorithms described here could be seen as the foundation of modern batch reinforcement learning, several other algorithms have been referred to as 'batch' or 'semi-batch' algorithms in the past. Furthermore, the borders between 'online', 'offline', 'semi-batch', and 'batch' can not be drawn distinctly; there are at least two different perspectives to look at the problem. Figure 4 proposes an ordering of online, semi-batch, growing batch, and batch reinforcement learning algorithms. On one side of the tree we have pure online algorithms like classic Q-

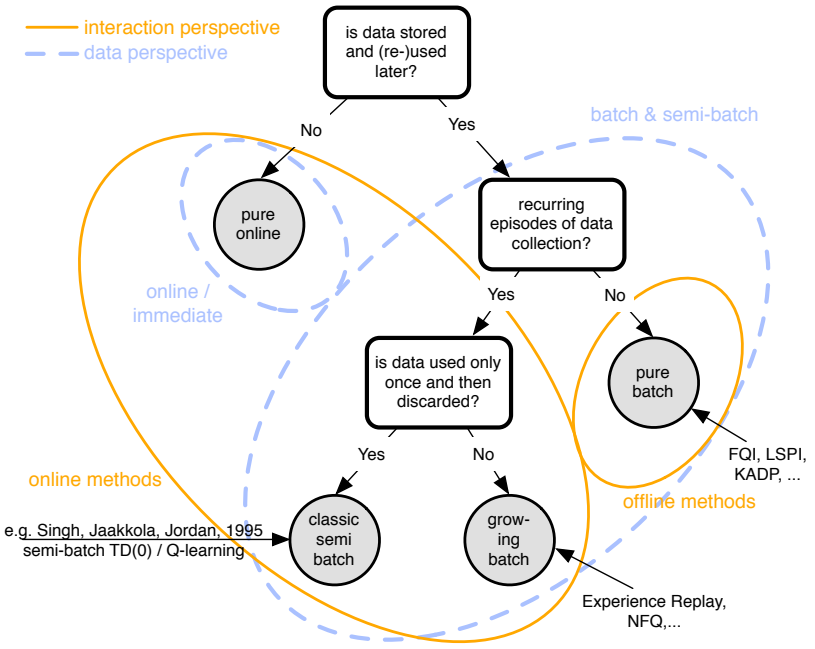

Fig. 4 Classification of batch vs. non-batch algorithms. With the interaction perspective and the data-usage perspective there are at least two different perspectives with which to define the category borders.

learning. On the opposite side of the tree we have pure batch algorithms that work completely 'offline' on a fixed set of transitions. In-between these extremal positions are a number of other algorithms that, depending on the perspective, could be classified as either online or (semi-)batch algorithms. For example, the growing batch approach could be classified as an online method—it interacts with the system like an online method and incrementally improves its policy as new experience becomes available—as well as, from a data usage perspective, being seen as a batch-algorithm, since it stores all experience and uses 'batch methods' to learn from these observations. Although FQI—like KADP and LSPI—has been proposed by Ernst as a pure batch algorithm working on a fixed set of samples, it can easily be adapted to the growing batch setting, as, for example, shown by Kalyanakrishnan and Stone (2007). This holds true for every 'pure' batch approach. On the other hand, NFQ (see section 6.1), which has been introduced in a growing-batch setting, can also be adapted to the pure batch setting in a straight-forward manner. Another class is formed by the 'semi-batch' algorithms that were introduced in the 90's (Singh et al, 1995) primarily for formal reasons. These algorithms make an aggregate update for several transitions—so it is not pure online learning with immediate updates. But what they do not do, however, is store and reuse the experience after making this update—so its not a full batch approach either.

5 Theory of Batch RL

The compelling feature of the batch RL approach is that it grants stable behavior for Q-learning-like update rules and a whole class of function approximators (averagers) in a broad number of systems, independent of a particular modeling or specific reward function. There are two aspects to discuss: a) stability, in the sense of guaranteed convergence to a solution and b) quality, in the sense of the distance of this solution to the true optimal value function.

Gordon (1995a,b) introduced the important notion of the 'averager' and proved convergence of his model-based fitted value iteration for this class of function approximation schemes by first showing their non-expansive properties (in maximum norm) and then relying on the classical contraction argument (Bertsekas and Tsitsiklis, 1996) for MDPs with discounted rewards. For non-discounted problems he identified a more restrictive class of compatible function approximators and proved convergence for the 'self-weighted' averagers (Gordon, 1995b, section 4). Ormoneit and Sen extended these proofs to the model-free case; their kernel-based approximators are equivalent to the 'averagers' introduced by Gordon (Ormoneit and Sen, 2002). Approximated values must be a weighted average of the samples, where all weights are positive and add up to one (see section 4.1). These requirements grant the non-expansive property in maximum norm. Ormoneit and Sen showed that their random dynamic programming operator using kernel-based approximation contracts the approximated function in maximum norm for any given set of samples and, thus, converges to a unique fixed point in the space of possible approximated functions. The proof has been carried out explicitly for MDPs with discounted rewards (Ormoneit and Sen, 2002) and average-cost problems (Ormoneit and Glynn, 2001, 2002).

Another important aspect is the quality of the solution found by the algorithms. Gordon gave an absolute upper bound on the distance of the fixed point of his fitted value iteration to the optimal value function (Gordon, 1995b). This bound depends mainly on the expressiveness of the function approximator and its 'compatibility' with the optimal value function to approximate. Apart from the function approximator, in model-free batch reinforcement learning the random sampling of the transitions obviously is another aspect that influences the quality of the solution. Therefore, for KADP, there is no absolute upper bound limiting the distance of the approximate solution given a particular function approximator. Ormoneit and Sen instead proved the stochastic *consistency* of their algorithm—actually, this could be seen as an even stronger statement. Continuously increasing the size of samples in the limit guarantees stochastic convergence to the optimal value function under certain assumptions (Ormoneit and Sen, 2002). These assumptions (Ormoneit and Sen, 2002, appendix A)—besides other constraints on the sampling of transitions—include smoothness constraints on the reward function (needs to be a Lipschitz continuous function of s, a and s') and the kernel. A particular kernel used throughout their experiments (Ormoneit and Glynn, 2001) that fulfills these constraints is derived from the 'mother kernel'

$$k_{\mathcal{F}_a,b}(s,\sigma) = \phi^+\left(\frac{||s-\sigma||}{b}\right) \Bigg/ \sum_{(s_i,a_i,r_i,s_i')\in\mathcal{F}_a} \phi^+\left(\frac{||s_i-\sigma||}{b}\right)$$

with ϕ^+ being a univariate Gaussian function. The parameter b controls the 'bandwidth' of the kernel—that is its region of influence, or, simply, its 'resolution'. Relying on such a kernel, the main idea of their consistency proof is to first define an 'admissible' reduction rate for the parameter b in dependence of the growing number of samples and then prove the stochastic convergence of the series of approximations $\hat{V}^k$ under this reduction rate to the optimal value function. Reducing the bandwidth parameter b can be interpreted as increasing the resolution of the approximator. When reducing b, the expected deviation of the implicitly-estimated transition model from the true transition probabilities—in the limit—vanishes to zero as more and more samples become available. It is important to note that increasing the resolution of the approximator is only guaranteed to improve the approximation for smooth reward functions and will not necessarily help in approximating step functions, for example—thus, the Lipschitz constraint on the reward function.

Besides these results, which are limited to the usage of averagers within the batch RL algorithms, there are promising new theoretical analysis of Antos, Munos, and Szepesvari, that presently do not cover more general function approximators, but, however, may lead to helpful results for non-averagers in the future (Antos et al, 2008).

6 Batch RL in Practice

In this section, we will present several applications of batch reinforcement learning methods to real-world problems.

6.1 Neural Fitted Q Iteration (NFQ)

The ability to approximate functions with high accuracy and to generalize well from few training examples makes neural networks—in particular, multi-layer perceptrons (Rumelhart et al, 1986; Werbos, 1974)—an attractive candidate to represent value functions. However, in the classical online reinforcement learning setting, the current update often has unforeseeable influence on the efforts taken so far. In contrast, batch RL changes the situation dramatically: by updating the value function simultaneously at all transitions seen so far, the effect of destroying previous efforts can be overcome. This was the driving idea behind the proposal of Neural Fitted Q Iteration (NFQ, Riedmiller (2005)). As a second important consequence, the simultaneous update at all training instances makes the application of batch supervised learning algorithms possible. In particular, within the NFQ framework, the adaptive

supervised learning algorithm Rprop (Riedmiller and Braun, 1993) is used as the core of the fitting step.

NFQ_main() {
input: a set $\mathscr{F}$ of transition samples (s,a,r,s') (same, as used throughout the text)
output: approximation $\hat{Q}^N$ of the Q-value function
    i=0
    init_MLP() $\rightarrow \hat{Q}^0$;
    Do {
        generate_pattern_set $P = \{(input_t; target_t), t = 1,\ldots,\#D\}$ where:
            $input_t = s,a,$
            $target_t = r + \gamma \max_{a' \in A} \hat{Q}^i(s',a')$
        add_artificial_patterns(P)
        normalize_target_values(P)
        scale_pattern_values(P)
        Rprop_training(P) $\rightarrow \hat{Q}^{i+1}$
        $i \leftarrow i+1$
    } While ($i < N$)

Fig. 5 Main loop of NFQ.

The implementation of the batch RL framework using neural networks is fairly straight-forward, as the algorithm in figure 5 shows. However, there are some additional tricks and techniques that help to overcome some of the problems that occur when approximating (Q-)value functions by multi-layer perceptrons:

- scaling input and target values is crucial for success and should always be done when using neural networks. A sensible scaling can be easily realized, since all training patterns are known at the beginning of training.
- adding artificial training patterns (also called 'hint-to-goal'-heuristic in Riedmiller (2005)). Since the neural network generalizes from collected experiences, it can be observed that the network output tends to increase to its maximum value if no or too few goal-state experiences with zero path costs are included in the pattern set. A simple method to overcome this problem is to build additional artificial (i.e. not observed) patterns within the goal region with target value zero that literally 'clamp' the neural network output in that region to 0. For many problems this method is highly effective and can be easily applied. When the target region is known, which is typically the case, no extra knowledge is used in applying this method.
- the 'Qmin-heuristic': normalizing' the 'Q' target values (Hafner and Riedmiller, 2011). A second method used to curtail the effect of increasing output values is to carry out a normalization step by subtracting the lowest target value from all target values. This results in a pattern set that has a target value of 0 for at least one training pattern. This method has the advantage in that no additional knowledge about states in the target regions need be known in advance.

Fig. 6 Brainstormers MidSize league robot. The difficulty of dribbling lies in the fact, that by the rules at most one third of the ball might be covered by the robot. Not loosing the ball while turning therefore requires a sophisticated control of the robot motion.

- using a smooth immediate cost-function (Hafner and Riedmiller, 2011). Since multi-layer perceptrons basically realize a smooth mapping from inputs to outputs, it is reasonable to also use a smooth immediate cost function. As an example, consider the immediate cost function that gives constant positive costs outside the target region and 0 costs inside the target region. This leads to a minimum time control behavior, which is favorable in many applications. However, accordingly, the path costs have rather crispy jumps, which are kind of difficult to represent by a neural network. Replacing this immediate cost function with a smoothed version, the main characteristic of the policy induced by the crisp immediate cost function is widely preserved while the value function approximation is much smoother. For more details, see Hafner and Riedmiller (2011).

6.2 NFQ in Control Applications

Applying reinforcement learning to the control of technical processes is particularly appealing, since it promises to autonomously learn optimal or near optimal controllers even in the presence of noise or nonlinearities without knowing process behavior in advance. The introduction of batch RL has contributed to a major breakthrough in this domain, since, due to its data efficiency, it is now possible to learn complex control behavior from scratch by directly interacting with the real system. Some recent examples of NFQ in real-world applications are learning to swing-up and balance a real cart-pole system, time optimal position control of pneumatic devices, and learning to accurately steer a real car within less than half an hour of driving (Riedmiller et al, 2007).

The following briefly describes the learning of a neural dribble controller for a RoboCup MidSize League robot (for more details, see also Riedmiller et al (2009)). The autonomous robot (figure 6) uses a camera as its main sensor and is fitted with an omnidirectional drive. The control interval is 33 ms. Each motor command con-

sists of three values denoting v_y^{target} (target forward speed relative to the coordinate system of the robot), v_x^{target} (target lateral speed) and v_θ^{target} (target rotation speed).

Dribbling means being able to keep the ball in front of the robot while turning to a given target. Because the rules of the MidSize league forbid simply grabbing the ball and only allow one-third of the ball to be covered by a dribbling device, this is quite a challenging task: the dribbling behavior must carefully control the robot, such that the ball does not get away from the robot when it changes direction.

The learning problem is modelled as a stochastic shortest path problem with both a terminal goal state and terminal failure states. Intermediate steps are punished by constant costs of 0.01.[2] NFQ is used as the core learning algorithm. The computation of the target value for the batch training set thus becomes:

$$\bar{q}_{s,a}^{i+1} := \begin{cases} 1.0 & , \quad \text{if } s' \in S^- \\ 0.01 & , \quad \text{if } s' \in S^+ \\ 0.01 + \min_{a' inA} \hat{Q}^i(s', a') & , \quad else \end{cases} \tag{12}$$

where S^- denotes the states at which the ball is lost, and S^+ denotes the states at which the robot has the ball and heads towards the target. State information contains speed of the robot in relative x and y direction, rotation speed, x and y ball position relative to the robot and, finally, the heading direction relative to the given target direction. A failure state $s \in S^-$ is encountered, if the ball's relative x coordinate is larger than 50 mm or less than -50 mm, or if the relative y coordinate exceeds 100 mm. A success state is reached whenever the absolute difference between the heading angle and the target angle is less than 5 degrees.

The robot is controlled by a three-dimensional action vector denoting target translational and rotational speeds. A total of 5 different action triples are used, $U = \{(2.0, 0.0, 2.0), (2.5, 0.0, 1.5), (2.5, 1.5, 1.5), (3.0, 1.0, 1.0), (3.0, -1.0, 1.0)\}$, where each triple denotes $(v_x^{target}, v_y^{target}, v_\theta^{target})$.

Input to the Neural Fitted Q Iteration method is a set of transition tuples of the form (state, action, cost, successor state) where the cost has either been 'observed' (external source) or is calculated 'on-the-fly' with the help of a known cost function $c : S \times A \times S \mapsto R$ (internal source). A common procedure used to sample these transitions is to alternate between training the Q-function and then sampling new transitions episode-wise by greedily exploiting the current Q-function. However, on the real robot this means that between each data collection phase one has to wait until the new Q-function has been trained. This can be aggravating, since putting the ball back on the play-field requires human interaction. Therefore, a batch-sampling method is used, which collects data over multiple trials without relearning.

The value function is represented by a multi-layer perceptron with 9 input units (6 state variables and 3 action variables), 2 hidden layers of 20 neurons each and 1 output neuron. After each batch of 12 trials, 10 NFQ iterations are performed. Learning the target values can be done within 300 epochs of supervised batch learning, using the Rprop learning method with standard parameters. After the learning

[2] Please note: costs = negative rewards. In this technical setting it's more natural to minimize costs, what, in principle, is equivalent to maximizing (negative) rewards.

is finished, the new controller can be used to control the robot during a subsequent data collection phase. In our experiments, after 11 batches (= 132 trials), a very good controller was learned. The complete learning procedure took about one and a half hours, including the time used for offline updating of the neural approximation of the Q-function. Including preparation phases, the actual interaction time with the real robot was about 30 minutes.

The neural dribbling skill performed significantly better than the previously used hand-coded and hand-tuned dribbling routine, particularly in terms of space and time needed to turn to the desired target direction (see figure 7). The neural dribbling skill has been successfully used in the Brainstormers competition team since 2007. Using it, the Brainstormers won the RoboCup world championship 2007 in Atlanta, USA and third place at the world championship in 2008 in Suzhou, China.

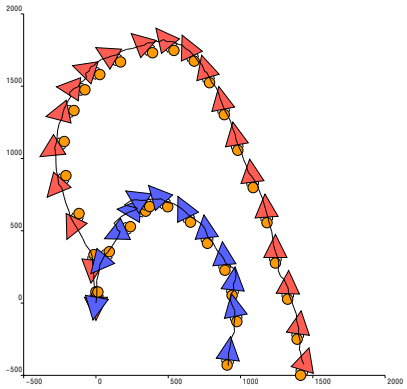

Fig. 7 Comparison of hand-coded (red) and neural dribbling behavior (blue) when requested to make a U-turn. The data was collected on the real robot. When the robot gets the ball, it typically has an initial speed of about 1.5 to 2 m/s in forward direction. The position of the robots are displayed every 120 ms. The U-turn performed by the neural dribbling controller is much sharper and faster.

6.3 Batch RL for Learning in Multi-Agent Systems

While the previous two sections have pointed to the advantages of combining the data-efficiency of batch-mode RL with neural network-based function approximation schemes, this section elaborates on the benefits of batch methods for cooperative multi-agent reinforcement learning. Assuming independently learning agents, it is obvious that those transitions experienced by one agent are strongly affected by the decisions concurrently made by other agents. This dependency of single transitions on external factors, i.e. on other agents' policies, gives rise to another argument for batch training: While a single transition tuple contains probably too little

information for performing a reliable update, a rather comprehensive batch of experience may contain sufficient information to apply value function-based RL in a multi-agent context.

The framework of decentralized Markov decision processes (DEC-(PO)MDP, see chapter **??** or Bernstein et al (2002)) is frequently used to address environments populated with independent agents that have access to local state information only and thus do not know about the full, global state. The agents are independent of one another both in terms of acting as well as learning. Finding optimal solutions to these types of problems is, generally, intractable, which is why a meaningful goal is to find approximate joint policies for the ensemble of agents using model-free reinforcement learning. To this end, a local state-action value function $Q_k : S_k \times A_k$ is defined for each agent k that it successively computes, improves, and then uses to choose its local actions.

In a straightforward approach, a batch RL algorithm (in the following, the focus is put on the use of NFQ) might be run independently by each of the learning agents, thus disregarding the possible existence of other agents and making no attempts to enforce coordination across them. This approach can be interpreted as an 'averaging projection' with Q-values of state-action pairs collected from both, cooperating and non-cooperating agents. As a consequence, the agents' local Q_k functions underestimate the optimal joint Q-function. The following briefly describes a batch RL-based approach to sidestep that problem and points to a practical application where the resulting multi-agent learning procedure has been employed successfully (for more details, see also Gabel and Riedmiller (2008b)).

For a better estimation of the Q_k values, the inter-agent coordination mechanism introduced in Lauer and Riedmiller (2000) can be used and integrated within the framework of Fitted Q Iteration. The basic idea here is that each agent always optimistically assumes that all other agents behave optimally (though they often will not, e.g. due to exploration). Updates to the value function and policy learned are only performed when an agent is certain that a superior joint action has been executed. The performance of that coordination scheme quickly degrades in the presence of noise, which is why determinism in the DEC-MDP's state transitions must be assumed during the phase of collecting transitions. However, this assumption can be dropped when applying the policies learned.

For the multi-agent case, step 1 of FQI (cf. equation (6)) is modified: Each agent k collects its own transition set $\mathcal{F}_k$ with local transitions (s_k, a_k, r_k, s'_k). It then creates a reduced (so-called 'optimistic') training pattern set $\mathcal{O}_k$ such that $|\mathcal{O}_k| \leq |P_k|$. Given a deterministic environment and the ability to reset the system to a specific initial state during data collection, the probability that agent k enters some s_k more than once is greater than zero. Hence, if a certain action $a_k \in A_k$ has been taken multiple times in s_k, it may—because of differing local actions selected by other agents—yield very different rewards and local successor states for k. Instead of considering all tuples from $\mathcal{F}_k$, only those that have resulted in maximal expected rewards are used for creating $\mathcal{O}_k$. This means that we assume that all other agents take their best possible local action, which is—when combined with a_k—most suitable for the current global state. Accordingly, the optimistic target Q-values $q^{i+1}_{s_k,a_k}$ for a given

local state-action pair (s_k, a_k) of agent k is computed according to

$$q^{i+1}_{s_k,a_k} := \max_{\substack{(s,a,r,s') \in \mathscr{F}_k, \\ s=s_k, a=a_k}} \left(r + \gamma \max_{a' \in A_k} \hat{Q}^i_k(s', a') \right).$$

Consequently, $\mathscr{O}_k$ realizes a partitioning of $\mathscr{F}_k$ with respect to identical values of s_k and a_k, and $q^{i+1}_{s_k,a_k}$ is the maximal sum of the immediate rewards and discounted expected rewards over all tuples $(s_k, a_k, \cdot, \cdot) \in \mathscr{F}_k$.

There are many applications that adhere to the problem class outlined, including production planning and resource allocation problems (Gabel and Riedmiller, 2008c). One particular class of problems that can be cast as decentralized MDPs is made up by job-shop scheduling problems (Brucker and Knust, 2005). Here, the goal is to allocate a specified number of jobs (also called tasks) to a limited number of resources (also called machines) in such a manner that some specific objective is optimized. Taking a batch reinforcement learning approach for scheduling, a learning agent is attached to each of the resources, receives local information about the set of jobs waiting for processing (features characterizing the jobs which are inputs to the neural networks used for value function approximation), and must decide which job to dispatch next. The agents interact repeatedly with the scheduling plant and they collect their own batches of transitions, which are then used to derive an improved dispatching policy by learning agent-specific state-action value functions $\hat{Q}_k$ using the NFQ adaptation outlined above. For well-established benchmark problems, the learned dispatching policies exhibit competitive performance, and, moreover, exhibit good generalization capabilities, meaning that they can be immediately applied to modified factory layouts (Gabel and Riedmiller, 2008a).

6.4 Deep Fitted Q Iteration

Present reinforcement learning algorithms, in general, are still limited to solving tasks with state spaces of rather low dimensionality. For example, learning policies directly from high-dimensional visual input—e.g. raw images as captured by a camera—is still far from being possible. Usually in such a task the engineer provides a method for extracting the relevant information from the high-dimensional inputs and for encoding it in a low-dimensional feature space of a feasible size. The learning algorithm is then applied to this manually constructed feature space.

Here, the introduction of batch reinforcement learning provides new opportunities for dealing directly with high-dimensional state spaces. Consider a set of transitions $\mathscr{F} = \{(s_t, a_t, r_{t+1}, s_{t+1}) | t = 1, \ldots, p\}$ where the states s are elements of a high-dimensional state space $s \in R^n$. The idea is to use an appropriate unsupervised learning method for learning a feature-extracting mapping from the data automatically. Ideally, the learned mapping $\phi : R^n \mapsto R^m$ with $m \ll n$ should encode all the 'relevant' information contained in a state s in the resulting feature vectors $z = \phi(s)$. The interesting thing now is, that by relying on the batch RL methods, we can com-

bine learning of feature spaces together with learning a policy within a stable and data-efficient algorithm. Within the growing batch approach, when starting a new learning phase, we would first use the data $\mathcal{F}$ to learn a new feature extraction mapping $\phi : R^n \mapsto R^m$ and then learn a policy in this feature space. This is done by first mapping all samples from the state space to the feature space, constructing a pattern set $\mathcal{F}_\phi\{(\phi(s),a,r,\phi(s'))|(s,a,r,s') \in \mathcal{F}\}$in the feature space and then applying a batch algorithm such as FQI. Since all experiences are stored in the growing batch approach, it is possible to change the mapping after each episode of exploration and improve it with the newly available data. A new approximation of the value function can be calculated immediately from the mapped transitions. Actually, using the set of transitions $\mathcal{F}$ it is possible to directly 'translate' an approximation $\hat{Q}^\phi$ that was calculated for the feature extraction ϕ to a new feature extraction ϕ' without loosing any information. We would simply apply one step of FQI with a slightly modified calculation of the target values:

$$\bar{q}_{\phi'(s),a} = r + \gamma \max_{a' \in A} \hat{Q}^\phi_{a'}(\phi(s')) \; .$$

When calculating the new target value for $\bar{q}_{\phi'(s),a}$ for the new feature vector $\phi'(s)$ of state s, this update uses the expected reward from the subsequent state s' as given by the old approximation $\hat{Q}^\phi$ using the feature vector $\phi(s')$ in the old feature space. These target values are then used to calculate a new approximation $\hat{Q}^{\phi'}$ for the new feature space.

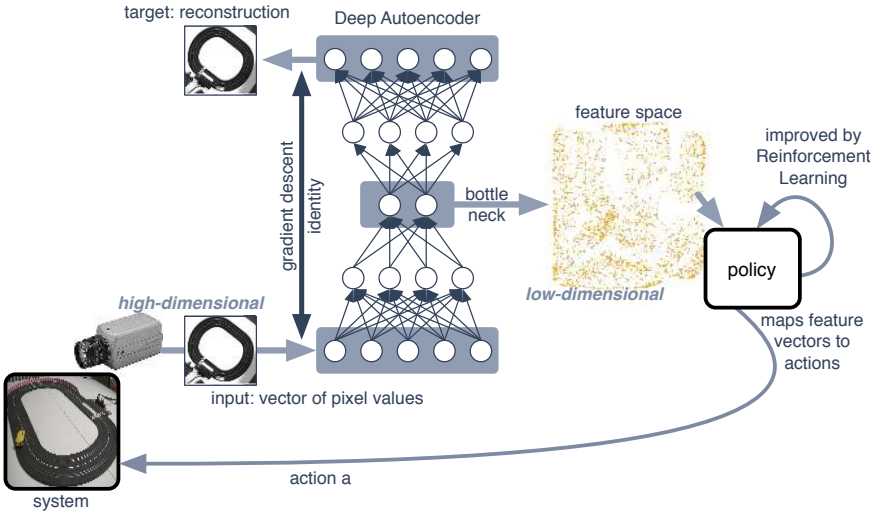

Fig. 8 Schematic drawing of Deep Fitted Q Iteration. Raw visual inputs from the system are fed into a deep auto-encoder neural network that learns to extract a low-dimensional encoding of the relevant information in its bottle-neck layer. The resulting feature vectors are then used to learn policies with the help of batch updates.

We have already implemented this idea in a new algorithm named 'Deep Fitted Q Iteration' (DFQ) (Lange and Riedmiller, 2010a,b). DFQ uses a deep auto-encoder neural network (Hinton and Salakhutdinov, 2006) with up to millions of weights for unsupervised learning of low-dimensional feature spaces from high dimensional visual inputs. Training of these neural networks is embedded in a growing batch reinforcement learning algorithm derived from Fitted Q Iteration, thus enabling learning of feasible feature spaces and useful control policies at the same time (see figure 8). By relying on kernel-based averagers for approximating the value function in the automatically constructed feature spaces, DFQ inherits the stable learning behavior from the batch methods. Extending the theoretical results of Ormoneit and Sen, the inner loop of DFQ could be shown to converge to a unique solution for any given set of samples of any MDP with discounted rewards (Lange, 2010).

The DFQ algorithm has been successfully applied to learning visual control policies in a grid-world benchmark problem—using synthesized (Lange and Riedmiller, 2010b) as well as screen-captured images (Lange and Riedmiller, 2010a)—and to controlling a slot-car racer only on the basis of the raw image data captured by a top-mounted camera (Lange, 2010).

6.5 Applications/ Further References

The following table provides a quick overview of recent applications of batch RL methods.

Domain	Description	Method	Reference
Technical process control, real world	Slot car racing	NFQ	Kietzmann and Riedmiller (2009)
Technical process control, real world	Dribbling soccer robot	NFQ	Riedmiller et al (2009)
Technical process control, real world	Steering an autonomous car	NFQ	Riedmiller et al (2007)
Technical process control, simulation	Nonlinear control benchmarks	NFQ, NFQCA	Hafner and Riedmiller (2011)
Technical process control, simulation	Pole swing up and balancing	Batch RL, Gaussian Processes	Deisenroth et al (2009)
Technical process control, simulation	Mobile wheeled pendulum	NFQ, FQI (Extra Trees)	Bonarini et al (2008)
Technical process control, simulation	Semi-active suspension control	Tree based batch RL	Tognetti et al (2009)
Technical process control, simulation	Control of a power system, comparison to MPC	FQI (Extra Trees)	Ernst et al (2005b), Ernst et al (2009)
Portfolio management, simulation	Managing financial transactions	KADP	Ormoneit and Glynn (2001)
Benchmarking, simulation	Mountain car, acrobot	FQI, CMAC	Timmer and Riedmiller (2007)
Multi agent systems, simulation	Decentralized scheduling policies	NFQ	Gabel and Riedmiller (2008a)
Multi agent systems, simulation	Keepaway soccer	FQI (NN, CMAC)	Kalyanakrishnan and Stone (2007)
Medical applications	Treatment of Epilepsy	Tree based batch RL	Guez et al (2008)

7 Summary

This chapter has reviewed both the historical roots and early algorithms as well as contemporary approaches and applications of batch-mode reinforcement learning. Research activity in this field has grown substantially in recent years, primarily due to the central merits of the batch approach, namely, its efficient use of collected data as well as the stability of the learning process caused by the separation of the dynamic programming and value function approximation steps. Besides this, various practical implementations and applications for real-world learning tasks have contributed to the increased interest in batch RL approaches.

References

Antos A, Munos R, Szepesvari C (2008) Fitted Q-iteration in continuous action-space MDPs. Advances in neural information processing systems 20:9–16

Baird L (1995) Residual algorithms: Reinforcement learning with function approximation. In: Proc. of the twelfth International Conference on Machine Learning, pp 30–37

Bernstein D, Givan D, Immerman N, Zilberstein S (2002) The Complexity of Decentralized Control of Markov Decision Processes. Mathematics of Operations Research 27(4):819–840

Bertsekas D, Tsitsiklis J (1996) Neuro-dynamic programming. Belmont, MA: Athena Scientific

Bonarini A, Caccia C, Lazaric A, Restelli M (2008) Batch reinforcement learning for controlling a mobile wheeled pendulum robot. In: IFIP AI, pp 151–160

Brucker P, Knust S (2005) Complex Scheduling. Springer, Berlin, Germany

Deisenroth MP, Rasmussen CE, Peters J (2009) Gaussian Process Dynamic Programming. Neurocomputing 72(7–9):1508–1524

Ernst D, Geurts P, Wehenkel L (2005a) Tree-Based Batch Mode Reinforcement Learning. Journal of Machine Learning Research 6(1):503–556

Ernst D, Glavic M, Geurts P, Wehenkel L (2005b) Approximate Value Iteration in the Reinforcement Learning Context. Application to Electrical Power System Control. International Journal of Emerging Electric Power Systems 3(1)

Ernst D, Glavic M, Capitanescu F, Wehenkel L (2009) Reinforcement learning versus model predictive control: a comparison on a power system problem. IEEE Transactions on Systems, Man, and Cybernetics, Part B 39(2):517–529

Gabel T, Riedmiller M (2008a) Adaptive Reactive Job-Shop Scheduling with Reinforcement Learning Agents. International Journal of Information Technology and Intelligent Computing 24(4)

Gabel T, Riedmiller M (2008b) Evaluation of Batch-Mode Reinforcement Learning Methods for Solving DEC-MDPs with Changing Action Sets. In: Proceedings of the 8th European Workshop on Reinforcement Learning (EWRL 2008), Springer, Lille, France, pp 82–95

Gabel T, Riedmiller M (2008c) Reinforcement Learning for DEC-MDPs with Changing Action Sets and Partially Ordered Dependencies. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), IFAAMAS, Estoril, Portugal, pp 1333–1336

Gordon GJ (1995a) Stable Function Approximation in Dynamic Programming. In: Proc. of the Twelfth International Conference on Machine Learning, Morgan Kaufmann, Tahoe City, USA, pp 261–268

Gordon GJ (1995b) Stable function approximation in dynamic programming. Tech. rep., CMU-CS-95-103, CMU School of Computer Science, Pittsburgh, PA

Gordon GJ (1996) Chattering in SARSA (λ). Tech. rep.

Guez A, Vincent RD, Avoli M, Pineau J (2008) Adaptive treatment of epilepsy via batch-mode reinforcement learning. In: AAAI, pp 1671–1678

Hafner R, Riedmiller M (2011) Reinforcement Learning in Feedback Control — challenges and benchmarks from technical process control. Machine Learning, accepted for publication, available online: DOI 10.1007/s10994-011-5235-x

Hinton G, Salakhutdinov R (2006) Reducing the Dimensionality of Data with Neural Networks. Science 313(5786):504–507

Kalyanakrishnan S, Stone P (2007) Batch reinforcement learning in a complex domain. In: The Sixth International Joint Conference on Autonomous Agents and Multiagent Systems, ACM, New York, NY, USA, pp 650–657

Kietzmann T, Riedmiller M (2009) The Neuro Slot Car Racer: Reinforcement Learning in a Real World Setting. In: Proceedings of the Int. Conference on Machine Learning Applications (ICMLA09), Springer, Miami, Florida

Lagoudakis M, Parr R (2001) Model-Free Least-Squares Policy Iteration. In: Advances in Neural Information Processing Systems 14, pp 1547–1554

Lagoudakis M, Parr R (2003) Least-Squares Policy Iteration. Journal of Machine Learning Research 4:1107–1149

Lange S (2010) Tiefes Reinforcement Lernen auf Basis visueller Wahrnehmungen. Dissertation, Universität Osnabrück

Lange S, Riedmiller M (2010a) Deep auto-encoder neural networks in reinforcement learning. In: International Joint Conference on Neural Networks (IJCNN 2010), Barcelona, Spain

Lange S, Riedmiller M (2010b) Deep learning of visual control policies. In: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2010), Brugge, Belgium

Lauer M, Riedmiller M (2000) An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems. In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Morgan Kaufmann, Stanford, USA, pp 535–542

Lin L (1992) Self-Improving Reactive Agents Based on Reinforcement Learning, Planning and Teaching. Machine Learning 8(3):293–321

Ormoneit D, Glynn P (2001) Kernel-based reinforcement learning in average-cost problems: An application to optimal portfolio choice. Advances in neural information processing systems 13 pp 1068–1074

Ormoneit D, Glynn P (2002) Kernel-based reinforcement learning in average-cost problems. IEEE Transactions on Automatic Control 47(10):1624–1636

Ormoneit D, Sen Ś (2002) Kernel-based reinforcement learning. Machine Learning 49(2):161–178

Riedmiller M (2005) Neural Fitted Q Iteration – First Experiences with a Data Efficient Neural Reinforcement Learning Method. In: Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Springer, Porto, Portugal, pp 317–328

Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Ruspini H (ed) Proceedings of the IEEE International Conference on Neural Networks (ICNN), San Francisco, pp 586 – 591

Riedmiller M, Montemerlo M, Dahlkamp H (2007) Learning to Drive in 20 Minutes. In: Proceedings of the FBIT 2007 conference., Springer, Jeju, Korea

Riedmiller M, Hafner R, Lange S, Lauer M (2008) Learning to dribble on a real robot by success and failure. In: Proc. of the IEEE International Conference on Robotics and Automation, pp 2207–2208

Riedmiller M, Gabel T, Hafner R, Lange S (2009) Reinforcement Learning for Robot Soccer. Autonomous Robots 27(1):55–74

Rumelhart D, Hinton G, Williams R (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536

Schoknecht R, Merke A (2003) Convergent combinations of reinforcement learning with linear function approximation. Advances in Neural Information Processing Systems 15 pp 1611–1618

Singh S, Jaakkola T, Jordan M (1995) Reinforcement learning with soft state aggregation. Advances in neural information processing systems 7 pp 361–368

Sutton R, Barto A (1998) Reinforcement Learning. An Introduction. MIT Press/A Bradford Book, Cambridge, USA

Timmer S, Riedmiller M (2007) Fitted Q Iteration with CMACs. In: Proceedings of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning (ADPRL 07), Honolulu, USA

Tognetti S, Savaresi S, Spelta C, Restelli M (2009) Batch reinforcement learning for semi-active suspension control. pp 582 –587

Werbos P (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University