VQQL: A Model to Generalize in Reinforcement Learning

Fernando Fernández, Daniel Borrajo, and Vicente Matellán

Universidad Carlos III de Madrid. Avda. de la Universidad, 30.
28911 Leganés, Madrid, Spain,
ffernand@grial.uc3m.es,dborrajo@ia.uc3m.es,vmo@ia.uc3m.es

Abstract Reinforcement learning has proven to be very successful for finding optimal policies on uncertain and/or dynamic domains. One of the problems on using such techniques appears with large state and action spaces. This problem appears very frequently given that most information in the type of tasks to which these techniques have been applied is continuous. In this paper, we describe a new mechanism for solving the states generalization problem in reinforcement learning algorithms, the VQQL technique.[1] This technique is based on the vector quantization technique for signal analog-to-digital conversion and compression, and on the Generalized Lloyd Algorithm for the design of vector quantizers. We show some results on applying this technique to learning skills for Robosoccer agents.

[1] It stands for *Vector Quantization for Q-Learning.*

1 Introduction

Real world for autonomous agents is dynamic and unpredictable. Thus, for most agent-based tasks, having a perfect domain theory (model) of how the actions of the agent affect the environment is usually an ideal. There are two ways of providing such models to agents planners/controllers: by careful and painstaking "ad-hoc" manual design of skills; or by automatically acquiring such skills. There have been already many different approaches for learning skills in such tasks (mainly robotic tasks), such as genetic algorithms for learning fuzzy rules [11], inductive collaborative techniques [5], or genetic algorithms for learning classifier systems [4, 12]. Among them, reinforcement learning techniques have proven to be very useful when modelling the robot worlds as MDP or POMDP problems [10, 17].

However, when using reinforcement learning techniques with large state and/or action spaces, an efficiency problem appears: the size of the state-action tables. Current solutions to this problem rely on applying generalization techniques to the states and/or actions. Some systems have used decision trees [2], neural networks [7], or variable resolution dynamic programming [13].

In this paper, we present an approach to solve the generalization problem that uses a numerical clustering method: the generalized Lloyd algorithm for the design of vector quantizers [8]. This technique is extensively employed for signal analog-to-digital conversion and compression, which have common characteristics to MDP problems. We have applied this technique for compacting the set of states that an agent perceives, thus dramatically reducing the reinforcement table size. We have used Q-learning [18] as the reinforcement learning technique, though we believe this can be made extensible to any other technique relying on a state-action table. In this paper, we have used the combination of vector quantization and reinforcement learning for acquiring the ball interception skill for agents playing in the Robosoccer simulator [15].

We introduce the reinforcement learning and the Q-learning algorithm in section 2. Then, the vector quantization technique and the generalized Lloyd algorithm are described in section 3. Section 4 describes how vector quantization is used to solve the generalization problem in the model VQQL, and in sections 5 and 6, the experiments performed to verify the utility of the model and the results are shown. Finally, the related work and conclusions are discussed.

2 Reinforcement Learning

The main objective of reinforcement learning is to automatically acquire knowledge to better decide what action an agent should perform at any moment to optimally achieve a goal. Among many different reinforcement learning techniques, Q-learning has been very widely used [18].

The Q-learning algorithm for deterministic Markov decision processes is described in table 1. It needs a definition of the possible states, $\mathcal{S}$, the actions that the agent can perform in the environment, $\mathcal{A}$, and the rewards that it receives at any moment for the states it arrives to after applying each action, r.

It dinamically generates a reinforcement table $Q(s,a)$ that allows it to follow a potentially optimal policy. Parameter γ controls the relative importance of past actions rewards with respect to new rewards.

Q-learning algorithm (S, A)

For each pair ($s \in \mathcal{S}$, $a \in \mathcal{A}$), initialize the table entry $Q(s,a)$ to 0.
Observe the current state s
Do forever

- Select an action a and execute it
- Receive immediate reward r
- Obseve the new state s'
- Update the table entry for $Q(s,a)$ as follows:

$$Q(s,a) \leftarrow r + \gamma \max_{a'} Q(s',a') \tag{1}$$

- Set s to s'

Table1. Q-learning algorithm.

For non-deterministic environments, the Q table update equation does not properly account for the probabilities involved in state transitions. Equation 2 solves the problem of environments in which the execution of the same action from the same state by an agent arrives to different states (different rewards could be obtained). In this case, parameter α refers to the probabilities involved, and it is computed using equation 3.

$$Q_n(s,a) \leftarrow (1-\alpha_n) Q_{n-1}(s,a) + \alpha_n \{r + \gamma \max_{a'} Q_{n-1}(s',a')\} \tag{2}$$

$$\alpha_n \leftarrow \frac{1}{1 + \text{visits}_n(s,a)} \tag{3}$$

where $\text{visits}_n(s,a)$ is the total number of times that the state-action entry has been visited.

3 Vector Quantization (*VQ*)

When dealing with digital communications, one of the main problems is how to convert analog signals, such as voice, into digital ones, without loosing quality. Also, when transmitting big amounts of digital information, it is necessary to compact that information efficiently. Vector quantization appeared as an appropriate way of solving these two problems, by reducing the number of bits needed to represent and transmit information [6].

In the case of large state spaces in reinforcement learning, the problem is similar: *how can we compactly represent a huge number of states with very few information?* In order to apply vector quantization to the reinforcement learning problem, we will provide first some definitions.

3.1 Definitions

Since our goal is to reduce the size of the reinforcement table, we have to find out a more compact representation of the states. If we have K attributes describing the states, and each attribute a_i can have *values*(a_i) different values, where this number is usually big (in most cases infinite, since they are represented with real numbers), then the number of potential states can be computed as:

$$S = \prod_{i=1}^{K} values(a_i) \tag{4}$$

Our goal is to reduce that number of states into N new states. These N states have to be able to approximately capture the same information as the S states; that is, all similar states in the previous representation, belong to the same new state in the new representation. The first definition takes this into account.

Definition 1. *A vector quantizer Q of dimension K and size N is a mapping from a vector (or a "point") in the K-dimensional Euclidean space, R^K, into a finite set C containing N output or reproduction points, called* code vectors, codewords, *or* codebook. *Thus,*

$$Q : R^K \longrightarrow C$$

where $C = (y_1, y_2, \ldots, y_N)$, $y_i \in R^K$.

This definition shows how an infinite set of vectors defined by R^K is translated to another finite one, C, of size N. Each new state has one value for each one of the old K attributes. This can be considered a clustering technique.

Given C (computed by the generalized Lloyd algorithm, explained below), and a vector $x \in R^K$, $Q(x)$ assigns x to the closest state from C. In order to define the closeness of one vector x to a state (vector) in C, we need to define a measure of the *quantization error*, which needs a *distortion measure* (analog to the similarity metric of clustering techniques).

Definition 2. *A distortion measure d is an assignment of a nonnegative cost $d(x, q)$ associated with quantizing any input vector $x \in R^K$ with a reproduction vector $q = Q(x) \in C$.*

In digital communications, the most convenient and widely used measure of distortion between an input vector x and a quantizer vector $q = Q(x)$, is the squared error or squared Euclidean distance between two vectors defined by equation 5.

$$d(x,q) = \|x,q\|^2 = \sum_{i=1}^{K}(x[i]-q[i])^2 \tag{5}$$

However, sometimes differences in one attribute value are more important than in another. In those cases, the weighted squared error measure is more useful, because it allows a different emphasis to be given to different vector components, as in equation 6. In other cases, the values $x[i]$ and $q[i]$ are normalized by the range of values of the attribute. This is a special case of the equation 6 where weights would be computed as the inverse of the square of the range (maximum possible value minus minimum possible value).

$$d(x,q) = \sum_{i=1}^{K} w_i(x[i]-q[i])^2 \tag{6}$$

Once defined a distortion measure, we can define Q as in equation 7.

$$Q(x) = \text{argmin}_{y \in C}\{d(x,y)\} \tag{7}$$

In order to measure the average error produced by quantizing M training vectors x_j with Q, average distortion is defined as the expected distortion calculated among any input vector and the quantizer Q:

$$D = \frac{1}{M}\sum_{j=1}^{M}\min_{y \in C}\{d(x_j,y)\} \tag{8}$$

Finally, we define *partition* and *centroid*, concepts needed for presenting the Lloyd algorithm for computing C from M input vectors.

Definition 3. *A partition or cell $R_i \subseteq R^K$ is the set of input vectors (old states) associated to the same (new) state in the codebook C.*

Definition 4. *We define the* centroid, *cent(R), of any set $R \subseteq R^K$ as that vector $y \in R^K$ that minimizes the distortion between any point x in R and y:*

$$cent(R) = \{y \in R^K \mid E[d(x,y)] \leq E[d(x,y')], \forall x \in R, y' \in R^K\} \tag{9}$$

where $E[z]$ is the expected value of z.

A common formula to calculate each component i of the centroid of a partition is given by equation 10.

$$cent(R)[i] = \frac{1}{\|R\|}\sum_{j=1}^{\|R\|} x_j[i] \tag{10}$$

where $x_j \in R$, $x_j[i]$ is the value of component (attribute) i of vector x_j, and $\|R\|$ is the cardinality of R.

3.2 Generalized Lloyd Algorithm (GLA)

The generalized Lloyd algorithm is a clustering technique, extension of the escalar case [9]. It consists of a number of iterations, each one recomputing the set of more appropriate partitions of the input states (vectors), and their centroids. The algorithm is shown in table 2. It takes as input a set T of M input states, and generates as output the set C of N new states (*quantization levels*).

Generalized Lloyd algorithm (T, N)

1. Begin with an initial codebook C_1.
2. Repeat
   (a) Given a codebook (set of clusters defined by their centroids) $C_m = \{y_i; i = 1, \ldots, N\}$, redistribute each vector (state) $x \in T$ into one of the clusters in C_m by selecting the one whose centroid is closer to x.
   (b) Recompute the centroids for each cluster just created, using the centroid definition in equation 10 to obtain the new codebook C_{m+1}.
   (c) If an empty cell (cluster) was generated in the previous step, an alternative code vector assignment is made (instead of the centroid computation).
   (d) Compute the average distortion for C_{m+1}, D_{m+1}
   Until the distortion has only changed by a small enough amount since last iteration.

Table2. The generalized Lloyd algorithm.

There are three design decisions to be made when using such technique:

Stopping criterion Usually, average distortion of codebook at cycle m, D_m, is computed and compared to a threshold θ $(0 \leq \theta \leq 1)$ as in equation 11.

$$(D_m - D_{m+1})/D_m < \theta \tag{11}$$

Empty cells One of the most used mechanisms consists of splitting other partitions, and reassigning the new partition to the empty one. All empty cells generated by the GLA are changed in each iteration by another cell. To define the new one, another non-empty cell with big average distortion y, is splitted in two:

$$y_1 = \{y[1] - \epsilon, \ldots, y[K] - \epsilon\}, \text{and}$$

$$y_2 = \{y[1] + \epsilon, \ldots, y[K] + \epsilon\}$$

Initial codebook generation We have used a version of the GLA as explained in table 3, that requires a partition split mechanism as the one described above inserted into the GLA in table 2.

GLA with Splitting (T)

1. Begin with an initial codebook C_1 with N (number of levels of the codebook) set to 1. The only level of the codebook is the centroid of the input.
2. Repeat
   (a) Set N to $N * 2$
   (b) Generate a new codebook C_{m+1} with N levels that includes the codebook C_m. The rest N undefined levels can be initialized to 0
   (c) Execute the GLA algorithm in table 2 with the splitting mechanism with parameters (T, N) over the codebook obtained in previous step

   Until N is the desired level

Table3. A version of the generalized Lloyd algorithm that solves the initial codebook and empty cell problems.

4 Aplication of *VQ* to *Q-learning.* VQQL

The use of vector quantization and the generalized Lloyd algorithm to solve the generalization problem in reinforcement learning algorithms requires two consecutive phases:

Learn the quantizer. Or to design the N-levels vector quantizer from input data obtained from the environment.

Learn the Q function. Once the vector quantizer is designed (we have clustered the environment in N different states), it is needed to learn the Q function, generating the Q table, that will be composed of N rows, and a column for each action (one could also use the same algorithm for quantizing actions).

We have two ways of unifying both phases:

Batch mode. We could obtain the information required to learn the quantizer and the Q function, and, later, learn both.

On-line mode. We could obtain data to generate only the vector quantizer, and, later, the Q function is learned by the interaction of the agent with the environment, using the previous designed quantizer.

The advantages of the first one are that it allows to use the same information several times, and the quantizer and the Q table are learnt with the same data. The second one allows the agent to use greedy strategies in order to increase the learning rate (exploration versus explotation).

In both cases, the behaviour of the agent, once the quantizer and the Q function are learnt, is the same; a loop that:

- Receives the current state, s, from the environment.
- Obtains the quantization level, s', or state to which the current state belongs.
- Obtains the action, a, from the Q table with bigger Q value for s'.
- Executes action a.

5 The Robosoccer domain

In order to verify the usefulness of the vector quantization technique to solve the generalization problem in reinforcement learning algorithms, we have selected a robotic soccer domain that presents us all the problems that we have defined in previous sections. The Robocup, and its Soccer Server Simulator, gives us the needed support.

The Soccer Server provides an environment to confront two teams of players (agents) [16]. Each agent perceives at any moment two types of information: visual and auditorial [16]. Visual information describes a player what it sees in the field. For example, an agent sees other agents, field marks such as the center of the field or the goals, and the ball. Auditorial information describes a player what it hears in the field. A player can hear messages from the referee, from its coach, or from other players. Any agent (player) can execute actions such as run (dash), turn (turn), send messages (say), kick the ball (kick), catch the ball (catch), etc.

One of the more basic skills a soccer player must have is ball interception. The importance of this skill comes from the dependency that other basic skills, such as kick or catch the ball, have with this one. Furthermore, ball interception is presented as one of the more difficult tasks to solve in the Robosoccer simulator, and it has been studied in depth by other authors [17]. In the case of Stone's work, neural networks were used to solve the ball interception problem with as a supervised learning task.

The essential difficulties of this skill come from the visual limitations of the agent, as well as from the noise that the simulator includes in movements of objects:

Visual Limitations The players can 'see' objects in limited areas defined by view cones, that are dependent of the view mode of the player. Furthermore, when an agent receives visual information of an object, this information is not exact, with a non-lineal noise, that is incremented with the distance to the objects.

Noise in Movements of Objects In order to reflect unexpected movements of objects in the real world, the Robosoccer simulator adds noise to the movement of objects and parameters of actions.

In order to intercept the ball, our agents parse the visual information that they receive from the simulator, and obtain the following information:[2]

- Relative Distance from the ball to the player.
- Relative Direction from the ball to the player.
- Distance Change, gives an idea of how Distance is changing.
- Direction Change, gives an idea of how Direction is changing.

[2] The Robosoccer simulator protocol version 4.21 has been used for training. In other versions of the simulator, other information could be obtained.

In order to intercept the ball, after knowing the values of these parameters, each player can execute several actions:

Turn changing the direction of the player according to a moment between -180 and 180 degrees.

Dash increasing the velocity of the player in the direction it is facing with a power between -30 and 100.

To reduce the number of possible actions that an agent can perform (generalization over actions problem), we have used macro-actions defined as follows. Macro-actions are composed of two consecutive actions: turn(T), and dash(D), resulting in turn-dash(T, D). We have selected $D = 100$, and T is computed according to $A + \Delta_A$, where A is the angle between the agent and the ball, and Δ_A can be: +45,+10,0,-10,-45. Therefore, we have reduced the set of actions to five actions.

To solve the state generalization problem, a single mechanism could be used, such as a typical scalar quantization on each parameter. In this case, the average quantization error, following the error quadratic distortion measure defined in equation 5, could be calculated as follows. The Distance parameter range is usually in (0.9,17.3). Thus, if we allow 0.5 as the maximum quantization error, we need near 17 levels. Direction is in the (-179,179) range. If we allow a quantization error of 2, 90 levels will be needed. Distance Change parameter is usually in (-1.9,1), so we need near 15 levels, allowing an error of 0.1, and Direction Change is usually in (-170,170), so we need 85 levels, allowing an error of 2. Then, following equation 4 we need $17 * 90 * 15 * 85 = 1,950,750$ states. This is an unabordable size for a reinforcement learning approach.

6 Results

In this section, the results of using the VQQL model for learning the ball interception skill in the Robosoccer domain are shown. In order to test the performance of the Lloyd algorithm, we generated a training set of 94.852 tuples (states) with the following iterative process, similar to the one used in [17]:

- The goalie starts at a distance of four meters in front of the center of the goal, facing directly away from the goal.
- The ball and the shooter are placed randomly at a distance between 15 and 25 from the defender.
- For each training example, the shooter kicks the ball towards the center of the goal with a maximum power (100), and an angle in the range $(-20, 20)$.
- The defender goal is to catch the ball. It waits until the ball is in a distance less or equal than 14, and then it starts to execute actions defined in section 5 while the goal is not in the catchable area [16]. Currently, we are only giving positive rewards. Therefore, if the ball is in the catchable area, the goalie tries to catch the ball, and if it succeeds, a positive reward is given to the last decision. If the goalie does not catch the ball, it can execute new actions.

Finally, if the shooter goals, or the ball goes out of the field, it receives a reward of 0.

Then, we used the algorithm described in Section 3 with different number of quantization levels (new states). Figure 1 shows the evolution of the average distortion of the training sequence. The x-axis shows the logarithm of the number of quantization levels, i.e. the number of different states what will be used afterwards by the reinforcement learning algorithm and the y-axis shows the average distortion obtained by GLA. The distortion measure used has been the quadratic error, as shown in equation 5. As it can be seen, when using 2^6 to 2^8 quantization levels, the distortion becomes practically 0.

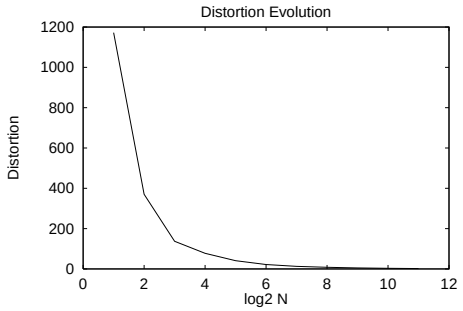


Figure1. Average distortion evolution depending on the number of states in the codebook.

In section 5, we computed the number of states needed for representing the environment for the ball interception task. It yielded $1,950,750$ states given a maximum quantization error of 0.5 (Distance), 2 (Direction), 0.1 (DistChng), and 2 (DirChng). The average distortion that is obtained according to equation 5 is:

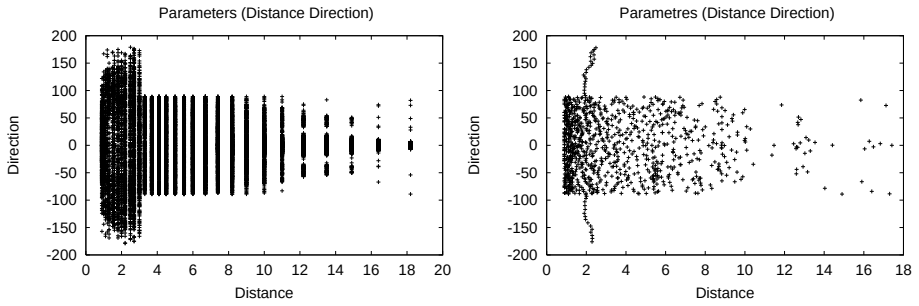
$$(\frac{0.5}{2})^2 + (\frac{2}{2})^2 + (\frac{0.1}{2})^2 + (\frac{2}{2})^2 = 2.7$$

given that the quantization error on each quantization is half of the maximum posible error. Instead, using the generalized Lloyd algorithm, with many less states, 2048,[3] the average distortion goes under 2.0. So, it reduces both the number of states to be represented, and the average quantization error.

Why is this reduction possible on the quantization error? The answer is given by the statistical advantages that the vector quantization provides over the scalar quantization. These advantages can be seen in Figure 2. In Figure 2(a), only the pairs of Distance and Direction that appeared in the training vectors have been plotted. As we can see, only some regions of the bidimensional space have

[3] In the following experiments we only provide results for 1024 quantization levels.

values, showing that not all combinations of the possible values of the Distance and Direction parameters exist in the training set of input states. Therefore, the reinforcement tables do not have to consider all possible combinations of these two parameters. Precisely, this is what vector quantization does. Figure 2(b) shows the points considered by 1024 states quantization. As it can be seen, it only generates states that represent minimally the states in the training set. The fact that there are parts of the space that are not covered by the quantization is due to the importance of the other two factors not considered in the figure (change in distance and direction).



(a) (b)

Figure2. Distance and Direction parameters from (a) original data, and (b) a codebook obtained by the GLA.

Up to this point, we have defined several possible quantizers or environment partitions, with different quantization errors, and we have to decide which of them to use in order to learning the skill. To answer this question and to obtain the better results in size of the Q table and performance of the goalie, we have learned a Q table per quantizer obtained before. We measure performance as the percentage of kicks of a new set of 100 testing problems that go towards the goal are catched by the goalie.

The results of these experiments are shown in Figure 3. In that figure the performance of the goalie is shown, depending of the size of the Q table. We show that with Q table sizes less than 128, a quasi-random behaviour is obtained. From sizes of the Q table from 128 to 1024, the performance increases until the maximum performance obtained up to now,[4] which is close to 60% of the plays. A very important comparison aspect to consider is that a random goalie would only achieve a 20% of success, and a goalie with the most used heuristic of *always go towards the ball* achieves only a 25% of successful behavior.

[4] We are still testing how bigger numbers of quantization levels affect the behavior.

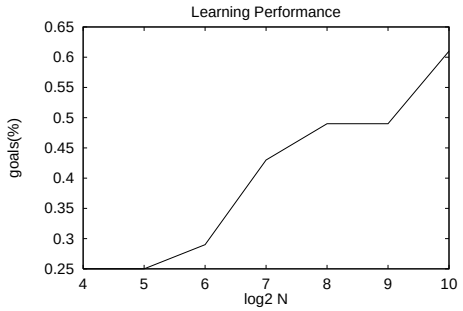


Figure3. Average distortion evolution depending on the number of states in the codebook.

7 Related Work

Other models to solve the generalization problem in reinforcement learning use decision trees as in the G-learning algorithm [2], and kd-trees (similar to a decision tree) in the VRDP algorithm [13]. Another solution is Moore's PartiGame algorithm [14] or neural networks [7]. One advantage of vector quantization is that it allows to easily define control parameters for obtaining different behaviors of the reinforcement learning technique. The main two parameters that have to be defined are number of quantization levels, and average distortion (similarity metrics). Other approaches to this problem were proposed in [3] and [1] where bayesian networks are used.

8 Conclusions and Future Work

We have shown how vector quantization and the generalized Lloyd algorithm allows us to dramatically reduce the number of states needed to represent a continuous environment. Furthermore, this technique gives us more quality in the quantization of these continuous environment than using a classical escalar quantization. The use of vector quantization for the generalization problem provides a solution to how to partition the environment into regions of states that can be considered the same for the purposes of learning and generating actions. It also solves the problem of knowing what granularity or placement of partitions is appropriate.

However, this mechanism introduce a set of open questions. As we explained above, the GL algorithm allows us to generate codebooks or sets of states of different sizes, each of them giving us different quantization errors. So, an important question is the relation between the number of quantization levels and the performance of the reinforcement learning algorithm. Another important issue relates to whether this technique can be applied not only to the state generalization problem, but also to actions generalization. We are also currently exploring the influence of providing negative rewards to the reinforcement learning technique.

References

1. Craig Boutilier, Richard Dearden, and Moises Goldszmidt. Exploiting structure in policy construction. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 1104–1111, Montreal, Quebec, Canada, August 1995. Morgan Kaufmann.
2. David Chapman and Leslie P. Kaelbling. Input generalization in delayed reinforcement learning: An algorithm and performance comparisions. *Proceedings of the International Joint Conference on Artificial Intelligence*, 1991.
3. Thomas Dean and Robert Givan. Model minimization in markov decision processes. In *Proceedings of the American Association of Artificial Intelligence (AAAI-97)*. AAAI Press, 1997.
4. Marco Dorigo. Message-based bucket brigade: An algorithm for the appointment of credit problem. In Yves Kodratoff, editor, *Machine Learning. European Workshop on Machine Learning*, LNAI 482, pages 235–244. Springer-Verlag, 1991.
5. Ramón García-Martínez and Daniel Borrajo. Learning in unknown environments by knowledge sharing. In John Demiris and Andreas Birk, editors, *Proceedings of the Seventh European Workshop on Learning Robots, EWLR'98*, pages 22–32, Edinburgh, Scotland, July 1998. University of Edinburgh Press.
6. Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression.* Kluwer Academic Publishers, 1992.
7. Long-Ji Lin. Scaling-up reinforcement learning for robot control. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 182–189, Amherst, MA, June 1993. Morgan Kaufman.
8. Yoseph Linde, André Buzo, and Robet M. Gray. An algorithm for vector quantizer design. In *IEEE Transactions on Communications, Vol1. Com-28, N° 1*, pages 84–95, 1980.
9. S. P. Lloyd. Least squares quantization in pcm. In *IEEE Transactions on Information Theory*, number 28 in IT, pages 127–135, March 1982.
10. S. Mahavedan and J. Connell. Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55:311–365, 1992.
11. Vicente Matellán, José Manuel Molina, and Camino Fernández. Genetic learning of fuzzy reactive controllers. *Robotics and Autonomous Systems*, 25(1-2):33–41, October 1998.
12. José M. Molina, Araceli Sanchis, Antonio Berlanga, and Pedro Isasi. An enhanced classifier system for autonomous robot navigation in dynamic environments. *Intelligent Automation and Soft Computing*, 1999. in press.
13. Andrew W. Moore. Variable resolution dynamic programming: Efficiently learning action maps in multivariate real-valued spaces. *Proceedings in Eighth International Machine Learning Workshop*, 1991.
14. Andrew W. Moore. The party-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, pages 711–718, San Mateo, CA, 1994. Morgan Kaufmann.
15. Itsuki Noda. Soccer server: A simulator of robocup. In *Proceedings of AI Symposium'95*. Japanese Society for Artificial Intelligence, December 1995.
16. Itsuki Noda. *Soccer Server Manual*, version 4.02 edition, January 1999.
17. Peter Stone. *Layered Learning in Multi-Agent Systems.* PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1999.
18. C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8(3/4):279–292, May 1992.