Data Mining for the category management in the retail market

Jochen Garcke, Michael Griebel and Michael Thess

January 29, 2009

1 Executive summary

Worldwide the retail market is under a severe competitive pressure. The retail trade in Germany in particular is internationally recognized as the most competitive market. To survive in this market most retailers use undirected mass marketing extensively. All prospective customers receive the same huge catalogues, countless advertising pamphlets, intrusive speaker announcements and flashy banner ads. In the end the customers are not only annoyed but the response rates of advertising campaigns are dropping for years. To avoid this, an individualization of mass marketing is recommended where customers receive individual offers specific to their needs. The objective is to offer the right customer at the right time for the right price the right product or content. This turns out to be primarily a mathematical problem concerning the areas of statistics, optimization, analysis and numerics. The arising problems of regression, clustering, and optimal control are typically of high dimensions and have huge amounts of data and therefore need new mathematical concepts and algorithms.

The underlying concept is the (semi)-automatic knowledge discovery via the analysis of huge databases, also known as data mining. The algorithmic core of the data mining process is called machine learning. This subject area originally belonged to computer science; the connection to statistics played a significant role from the beginnings. In recent years, further mathematical aspects were being considered especially in research, an example is the field of statistical learning theory. Algorithms with such a background are used successfully in many applications, not least due to their mathematical foundation.

The underlying assumption is that similar customer data signifies similar customer behaviour, this allows to assess new customers on the basis of the behaviour of former customers. Of fundamental importance is that many modern machine learning approaches use the representation of functions over high dimensional attribute spaces. This allows the coupled non-linear treatment of different attributes like income, debt, number of children, or type of car which results in an improved estimation of the likely customer behaviour. Approximation theory and numerics already play a substantial role for the development

of new and the improvement of existing machine learning approaches. This will intensify in the future.

Nowadays the new numerical approaches for moderate high dimensional problems are advanced far enough that they can be used in first real live data mining applications for the category management in the retail market. Further research to extend the approaches to really high dimensions is necessary to allow the efficient treatment of even more attributes. Focused research programs should be setup to accelerate this development. With such additional support the development of new data mining methods could be pushed sufficiently far that even fully automatic interactive systems for the category management could be brought into maturity. This would help to strengthen the international competitiveness of the German retail sector.

This article describes the role of mathematics for the category management in trade, in particular the role of approximation theory and numerics. Current state of art and success stories are outlined, and new developments and challenges are given.

2 Category management in the retail market: overview and status quo

The retail market is an especially dynamic one. This is traditionally due to the similarity in the offered products since all retailers have access to more or less the same range of products via their distributors. In the last years the internet allowed new business concepts and further intensified internationalization and increased competitive pressure. For the application of a typical data mining process many, mostly anonymous data of the customer behaviour is available, which can be used for the optimization of the offers.

The problems arising in category management can be separated into four different areas:

- campaign optimization (i. e., selection of target groups and customers),
- cross- and up-selling (i. e., additional sales to customers),
- assortment optimization (i. e., product assortment and categories),
- price optimization (i. e., optimization of product prices and promotions).

A few years ago retailers started to apply mathematical approaches for the analysis of customer behaviour, but its use is sporadic and differs according to the line of business and the marketing activity. While the e-business shows a remarkable adoption of, arguably often too simple, algorithms, the mail order business uses mathematics to a large degree only for the optimization of mailing activities, that is the selection of customers with a high response probability to special offers. Last is the stationary retail market, but exactly here the current technological revolution of interactive digital shopping devices opens

up new interesting possibilities for the development and application of new mathematical methods for the category management.

The high degree of customer interaction in retail is beneficial for the use of mathematical approaches since a large amount of customer data is available. At first the use of mathematics proved useful in some classical data mining fields. Here, it is very common to use classification algorithms for the optimization of mailings. Clustering methods for the segmentation of customers into thematic groups are increasingly successful. Other areas like real-time analysis and offers only apply the simplest methods. In the strategic field of management of commodity groups, that is the optimization of the range of products and their prices, the use of modern mathematical instruments is still the exception. But exactly here is, in combination with real-time approaches of optimal control, an important upcoming application area for the interdisciplinary cooperation of business, computer science and mathematics. Further information on data mining approaches in retail, marketing and customer relationship management can for example be found in [2].

In the following, we focus on the first two problems, i. e., campaign optimization and cross- and up-selling, and existing success stories of the use mathematical approaches in these fields.

2.1 Optimization of campaigns

With regard to the use of mathematics the optimization of campaigns is the most advanced. The goal is to apply marketing campaigns with a clear focus on the target customers. This concerns both, the definition of the aims and procedure of the campaign, as well as the analysis of the results. One distinguishes here between target group (segmentation) and target customer (individualization). While target groups are strictly defined according to one or several attributes (for example female), target customers are selected based on an individual assessment in form of a numerical value, the score. An example for segmentation is the mailing of a catalogue of sporting goods only to customers interested in sports, i. e., those who bought sporting goods before. For the individualization on the other hand each customer is checked for affinity to this specific catalogue of sporting goods, independent of being part of the segment of sport affine customers. For segmentation mostly clustering algorithms are used, while for the individualization mainly classification and regression algorithms are applied. In the following we will discuss the case of mailing optimization in more detail.

Success story: Optimization of mailings For the mailing of catalogues mostly still all customers in the address list are selected, independent of their response probability. This is called unpersonalized mailing. The costs of such a mailing campaign consists of fixed costs (primarily creation and printing of the catalogue), the shipping costs (S), the follow up (FU), and the order processing (OP). Tab. 1 shows the calculation of the profit for an exemplary mailing campaign with 100.000 recipients, 1 % response probability and 500 EUR income

Table 1: Profit calculation for a classical mailing campaign

fixed costs	50.000 EUR	=	50.000 EUR
costs for S	100.000 · 1,50 EUR	=	150.000 EUR
costs for FU	98.500 · 1,50 EUR	=	147.750 EUR
costs for OP	2.500 · 5 EUR	=	12.500 EUR
total costs		=	360.250 EUR
income	1.000 · 500 EUR	=	500.000 EUR
profit		=	139.750 EUR

break-even-point: 721 responders

Table 2: Profit calculation for an optimized mailing campaign

fixed costs	50.000 EUR	=	50.000 EUR
costs for S	40.000 · 1,50 EUR	=	60.000 EUR
costs for FU	38.600 · 1,50 EUR	=	57.900 EUR
costs for PP	2.100 · 5 EUR	=	10.500 EUR
total costs		=	178.400 EUR
income	950 · 500 EUR	=	475.000 EUR
profit		=	296.600 EUR

break-even-point: already after 357 responders!

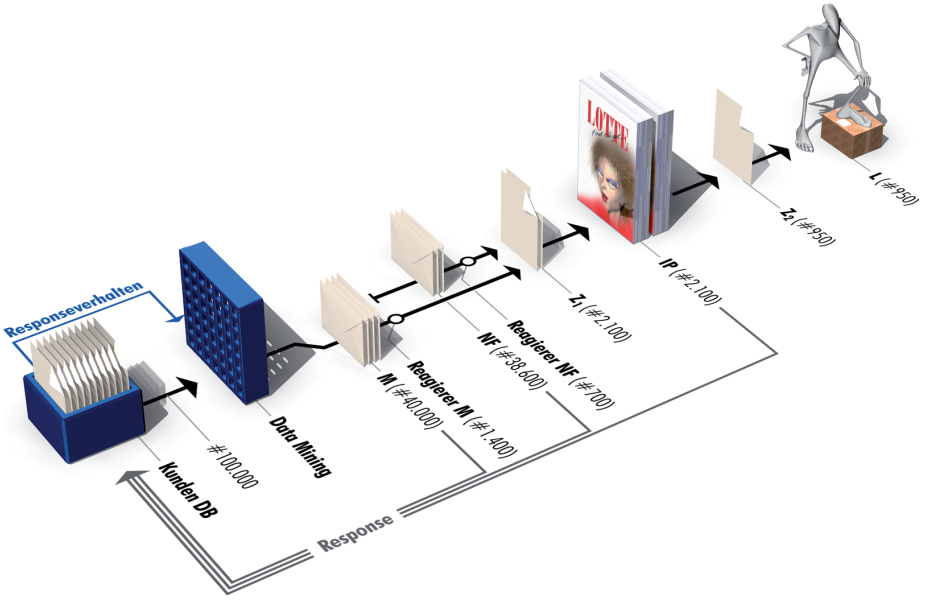
per responder. For simplification the revenue is considered profit.

To select only the customers with the highest response probability by means of data mining can increase the overall profit. The phases of such an optimized campaign are sketched in Fig. 1. The historical mailings for a catalogue are analysed based on the existing customer data. Models for the response probability are machine-learned based on the customer profile and then are evaluated on the 100.000 prospective recipients. This results in a list of scores out of which the 40.000 with the highest score (that is with the estimated largest response probability) are selected and sent the catalogue. If done correctly such a data mining approach can achieve more than twice the conventional response rate, here we assume 950 responders (ca. 2.4 %).

The calculation of profit in Tab. 2 shows that, in comparison to the classical mailing campaign, in the end a higher profit of ca. 297.000 EUR instead of ca. 140.000 EUR is generated with less income. As an additional benefit customers overall receive less catalogues. Such personalized mailing campaigns are successfully used by several mail order companies.

The optimization of a mailing campaign is based on the classification of customers using a model which is learned from existing customer data. An example for such a classification method is described in Fig. 2. Mathematically

Figure 1: The phases of an optimized mailing campaign



it is based on an approximate reconstruction and evaluation of functions over a high dimensional state space of customer attributes. Here, the method of sparse grids [5, 11] can be used for the approximation of such high dimensional functions. Alternatives are for example kernel based approaches with radial basis functions or neural networks. Many more classification algorithms exists in the literature, see [8] and the references cited therein.

2.2 Cross- and up-selling

As a second example for category management in the retail market we consider cross- and up-selling. Every salesperson knows that it is easier to sell additional products to an existing customer than to gain new customers. Cross- and up-selling addresses this core topic of increasing the customer value. The goal is to offer additional products (cross-selling) or higher valued products (up-selling) to existing customers based on their preferences which are indicated by their interests or former purchases. Besides the increase of revenue for the merchant good cross- and up-selling also leads to higher satisfaction of the customer. Since the customer is receiving offers he is actually interested in he can save time and can avoid searching on his own.

Cross-selling starts with the disposition of the products into the market. This is traditionally the role of the category manager, although mathematical approaches are being used for several years as well. In particular, clustering approaches are used for basket analysis. These methods work transaction based

Figure 2: Classification with sparse grids for the optimization of mailing campaigns

Classification and regression with regularization networks.
The optimization of a mailing campaign poses a *classification problem* which mathematically can be formulated as a *high dimensional approximation* problem.
The set $T = \{x_i \in \mathfrak{R}^d\}_{i=1}^M$ consists of M customers, who were recipients of a former mailing campaign and which are characterized by d attributes like sex, age, or profession. Furthermore, for each customer the target value $y_i \in \{-1, +1\}$ is known, which indicates if the customer responded to the campaign or not. The goal is to reconstruct a function from the given data which describes the likely relationship between attributes and the class label and therefore allows the prediction of the likelihood of a response for new customers. An example of such a function in two dimensions is shown in the adjoining figure.
The solution process consists of two phases. In the training phase the classifier is computed using the historical data, which describes the relationship between the customer attributes and the response probability. In the evaluation phase the learned model is evaluated for new customers.
To compute the classifier f a minimization problem

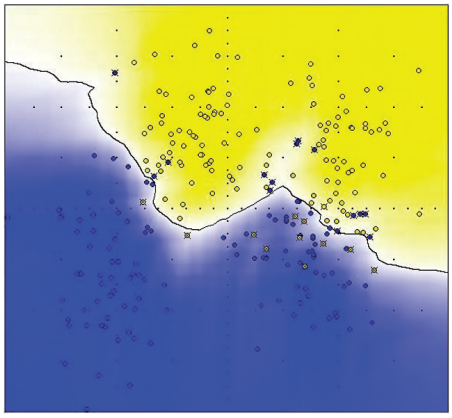
$$\min_{f \in V} R(f)$$

is solved in a suitable function space V, here R is an operator from V to $\mathfrak{R}$. For the most advanced current methods this can be written using the so-called regularization network approach [7]

$$R(f) = \frac{1}{M} \sum_{i=1}^{M} C\big(f(x_i, y_i)\big) + \lambda \phi(f)$$

where $C(x, y)$ is a cost functional measuring the distance between the given data and the classifier, e.g., $C(x, y) = (x - y)^2$, $\phi(f)$ is the regularization operator, which describes a-priori assumptions on f, e.g., $\phi(f) = ||\nabla f||^2$, and λ is a regularization parameter which balances these two terms.

Classification with a sparse grid function



In the actual numerical computation a discrete approximation of f is obtained. To cope with the high number of dimensions for example the sparse grid method can be employed [5, 11]. This approach scales only linearly with the number of data, for details see [6]. The method of sparse grids breaks the curse of dimension of discetization approaches based on conventional grids due to its use of a truncated tensor product multi-scale basis. Here, the number of employed grid points still grows exponentially with the number of dimensions but with a much smaller basis.
In the evaluation phase the classifier f is applied on the data of new customers. The higher the value of f, the so-called score, the higher the likelihood of response. Therefore, the customers are sorted according to their estimated score in descending order and the first n customers are chosen for the campaign. The choice of a suitable n depends on the cost of the campaign and the expected revenue per customer.
Nowadays up to several million of customers are analysed for mailing campaigns to achieve a selection as good as possible. The use of classification approaches achieves up to 100 % higher returns. Note that response rates for traditional mailings are typically below 1 %.

and analyse for example cashier data with regard to cooperative sales. Products which are frequently bought together can therefore easily be placed near to each other in the store. Alternatively content based methods are used to analyse products and categories according to their attributes (colour, description, sound, ...) and appropriate product clusters are formed.

In addition to the disposition of products the e-business brought new forms of interactive and automated cross-selling: recommendation engines and avatars lead the customer to related products and services. Well known and at the forefront is the online shop Amazon.com. While they are shopping customers are presented with overviews of related products based on their current shopping basket and product searches ("customers who bought this product also bought ..."). Although the early algorithms of Amazon.com were based on simple correlation analysis, they led the way for modern recommendation engines and adaptive analysis systems. Recommendation engines are nowadays established in e-business and are used in generalized forms for a wide range of applications like searches, matching, personalized pages, and dynamic navigation. At the same time it became the topic of academic research and meanwhile a large amount of publications exists.

Current methods range from clustering and text mining, Bayesian nets and neural nets up to complex hybrid solutions. Although the mathematical foundation of many approaches is still lacking, there is no doubt that currently an exciting research topic for applied mathematics is built here. In the following we discuss this example of dynamic programming for product recommendations in more detail.

Success story: Product recommendations Recommendation engines nowadays play an important role for automated customer interaction. A recommendation engine offers, based on click and purchase behaviour of a customer, automatically related product recommendations. The recommendation engine learns online directly from the customer interaction. Recommendation engines increase the sales up to 20 % and lead to enlarged customer satisfaction. But their application is not limited to this, modern recommendation engines vary design, product assortment and prices dependent on the user and allow totally new possibilities of personalization.

In stationary retail the use of automatic recommendation engines appeared until now technically infeasible, although interest exists since most buying decisions take place in the store. But change is on the horizon. In the first shopping malls electronic tools like the personal shopping assistant are available, a device which is placed on the shopping cart. Customers can access detailed information for a product from the shelf by using the scanner of the personal shopping assistant, the display then shows the corresponding information and additionally related product recommendations. Such systems allow for the first time fully automatic interaction with the customer in the store, for example in form of real time couponing on the receipt depending on the purchases or in form of dynamic price changes using electronic displays. This results in an interest-

Figure 3: Customer interaction using stochastic optimal control

Product recommendations as a reinforcement learning problem

The problems of adaptive product recommendations can be posed as a reinforcement learning problem which is mathematically based on the theory of optimal control. Besides theoretical studies already successfully commercial implementations exist. In reinforcement learning one considers a set of states s. For each state an action a from a set of actions can be chosen which leads to a new state s' with a scalar reward r. The sum of all rewards during a certain episode is to be maximized. In the simple case of a recommendation engine the states are the observed products, the actions are the recommended products and the reward corresponds to the price of a product in the event of a purchase. As long as the so-called Markov-property holds one can formulate a Markov decision process which is described by a discrete Bellman-equation:

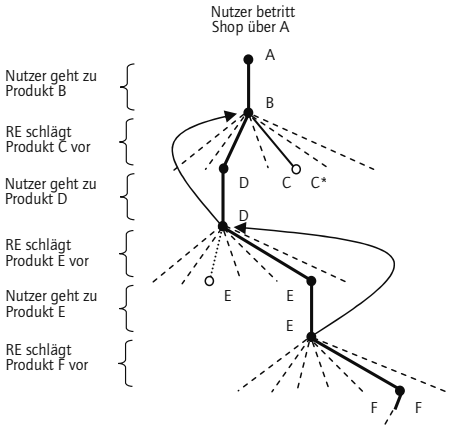
$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{s'} P^{a}_{ss'} \left[R^{a}_{ss'} + \gamma V^{\pi}(s') \right]$$

with

$P^{a}_{ss'}$	transaction probabilities
$R^{a}_{ss'}$	transition reward
γ	discount rate.

Here $V^{\pi}(s)$ is the state value function, which assigns the expected cumulative discounted reward over the remaining episode to each state s. One now searches for the optimal policy $\pi(s,a)^*$, which describes the stochastic choice of the best action a at state s. There $\pi(s,a)^*$ gives the optimal recommendation.

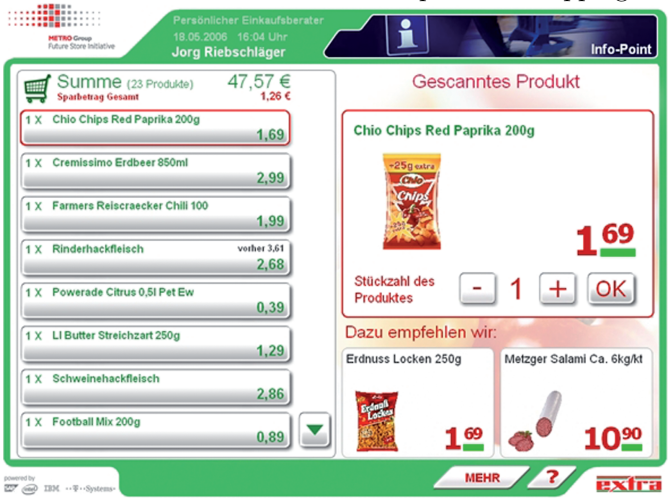
Interpretation of customer interaction as an interplay between recommendation engine and customer



Generalized methods from optimal control are used, in particular policy iteration and value iteration, to compute the optimal policy.

In many applications the state value function $V^{\pi}(s)$ can not be described in tabular form due to the large number of states. In these cases it is approximated with regression approaches. Again, the efficient and approximate representation of high dimensional functions plays a significant role. Furthermore one has to distinguish between reinforcement learning approaches which know a model and use it, that is which know the transition probabilities $P^{a}_{ss'}$ (e.g. via regression). Here classical methods from optimal control can be used which allow the computation for several million of states in a few hours. On the other hand there are model-free approaches – in particular Monte Carlo and TD-methods – which can be used for online learning.

Figure 4: Product recommendations in a personal shopping assistant



ing and broad new field for automatic analysis systems and recommendation engines for the generation of product recommendations and for price optimization.

A modern mathematical formulation of such adaptive product recommendations can be achieved with tools from stochastic optimal control and is discussed in Fig. 3. Here, a reinforcement learning formulation is used which leads to a discrete Hamilton-Jacobi-Bellman-equation. Again high dimensional state spaces arise for whose treatment often neural networks, kernel based approaches or decision trees are used [10]. The method of sparse grids can be applied here as well [9].

3 Outlook

An important aspect of successful data mining applications in retail is the approximate and efficient representation of high dimensional functions. Especially in the last decade significant improvements have been made to break the so-called curse of dimension [1], that is to develop and analyse methods whose complexity does not scale exponentially with the dimension of the underlying space. In addition to the sparse grid method [5] let us mention low-rank tensor product approaches [3, 4] as well as nonlinear approximation approaches like neural networks, kernel methods and LASSO [8]. Each technique has its own specific assumptions on the function and the data to be handled such that the curse of dimension can be avoided. Nevertheless, further research on numerical methods for high dimensional problems is absolutely necessary to improve the efficiency of the algorithms and to better understand their application possibil-

ities.

A significant new perspective for the presented category management in the retail market is the management of the full customer life cycle. Until now, marketing activities like campaigns, cross- and up-selling as well as product assortment and price optimization are treated independently. Often such measures to increase the revenue are not sustainable. In an ideal situation an optimization over all five dimensions (customer, content, time, channel, price) of the whole problem should take place instead to maximize the customer value over the whole customer life cycle.

A mathematical contribution can contain a quantification and optimization of the value of each customer. A revenue calculation per customer would be possible which then would be maximized using stochastic optimal control or reinforcement learning. Such extended optimization problems require the representation of functions (the customer value) over a high-dimensional space (the state space). Here, the state space represents the customer, the product, the price as well as the time and form of marketing activities. A company can take actions and place suitable ads at given times for suitable products with attractive prices. Of special interest for the future are the development of new adaptive regression algorithms, for example to optimize campaigns in real time, and optimal control approaches, for example to optimize the revenue of the whole customer life cycle. Both these problems involve high dimensional approximation problems.

Finally, some critical remarks on the commercial use of personal data and sociodemographic information are in order. Banks, insurance companies, and, as shown, increasingly the retail industry use customer profiles and risk groups for decision making. Good conditions are only given to good customers. Growing databases, loyalty cards, and the trade of customer data increases the trend. Studies warn of an increasing across the board assessment of customer groups. There is the danger that some people have no access to certain services or products due to companies who exempt them according to a data mining model which is based only on the available data.

Recent examples are customers who have to pay higher credit interest rates based on their place of living, insurants who can not get an occupational disablement insurance only due to some unspecified former illnesses, or people who were not able to rent an apartment due to late payments of mobile phone bills and a resulting bad credit entry. To avoid such undesired developments a highly transparent data mining solution with explanations is necessary (why did this recommendation happen). Existing laws and practices on data privacy protection need to be critically revisited under these aspect and might need to be modified.

4 Visions and suggested actions

For the future it is quite possible that data mining approaches like analytic systems and recommendation engines for product and price optimization will be

commonly used for a fully automatic interaction with customers in stores. The focus will be on interactivity and adaptivity, online learning will play an important role. Another trend is surely the management of the customer life cycle and the maximization of the customer life time value. To a large degree this would technically be possible nowadays, all data for the customer life cycle are in principle present somewhere in the company, but since they are not fully integrated in the IT-process they are typically not used this way. But the technical requirements for the use of data mining methods for the category management in the retail market are fulfilled. What needs to be done is a mathematical investigation of interactive and adaptive online learning methods to further their development. Until now mainly heuristic ad-hoc approaches exist. A mathematical formulation of adaptive product recommendations can be obtained with approaches from the field of stochastic optimal control via a reinforcement learning formulation which leads to a discrete Hamilton-Jacobi-Belmann-equation. To solve this equation efficiently and fast new algorithms and methods especially for high dimensional problems need to be developed, and existing methods need to be substantially improved. This can only be achieved in an interdisciplinary cooperation of mathematicians, computer scientists and end users. For this task specific funding programs need to be set up who allow interested groups to be active in this field. In the priority programme 1324 "Mathematical methods for the extraction of quantifiable information from complex systems" of the Deutsche Forschungsgemeinschaft as well as in the programme "Mathematics for innovations in industry and services" of the German Bundesministerium für Bildung und Forschung first important steps are undertaken in practical as well as theoretical aspects. This needs to be continued and extended into the future. Besides institutionalized funding financial support from industry and retail is necessary, which until now happened infrequently and quite risk averse.

References

[1] R. Bellman, Dynamic Programming, Princeton Univ. Press, 1957.

[2] Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Wiley and Sons, 2004.

[3] G. Beylkin and M. J. Mohlenkamp, Algorithms for Numerical Analysis in High Dimensions, SIAM J. Sci. Comput., 26:2133–2159, 2005.

[4] S. Börm, L. Grasedyck and W. Hackbusch, Hierarchical Matrices, Lecture Note 21, Max Planck Institute for Mathematics in the Sciences, Leipzig, 2003.

[5] H.J. Bungartz and M. Griebel, Sparse Grids, Acta Numerica, 13:147–269, 2004.

[6] J. Garcke, M. Griebel and M. Thess, Data Mining with Sparse Grids, Computing, 67:225–253, 2001.

[7] F. Girosi, M. Jones and T. Poggio, Regularization Theory and Neural Network Architectures, Neural Computation, 7:219–265, 1995.

[8] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, 2001.

[9] R. Munos, A Study of Reinforcement Learning in the Continuous Case by the Means of Viscosity Solutions, Machine Learning, 40(3):265–299, 2000.

[10] R.S. Sutton and A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.

[11] C. Zenger, Sparse Grids, In Parallel Algorithms for Partial Differential Equations, Proceedings of the Sixth GAMM-Seminar, Kiel, 1990, volume 31, Notes on Num. Fluid Mech., Vieweg-Verlag, 241–251, 1991.