ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ

ΚΩΝΣΤΑΝΤΙΝΟΣ Σ. ΔΑΛΑΜΑΓΚΙΔΗΣ
ΧΗΜΙΚΟΣ ΜΗΧΑΝΙΚΟΣ

ΕΦΑΡΜΟΓΗ ΕΝΙΣΧΥΜΕΝΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΑΝΕΣΗ ΚΑΙ ΕΞΟΙΚΟΝΟΜΗΣΗ ΕΝΕΡΓΕΙΑΣ ΣΕ ΚΤΙΡΙΑ

REINFORCEMENT LEARNING FOR ENERGY CONSERVATION AND COMFORT IN BUILDINGS

ΕΡΓΑΣΙΑ ΓΙΑ ΤΗΝ ΑΠΟΚΤΗΣΗ
ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

ΧΑΝΙΑ 2003

CONTENTS

1 INTRODUCTION

1.1 The problem of energy conservation and occupant comfort in buildings

During the 1960s and 1970s people became sensitive to the issue of energy conservation. The last 40 years many studies have investigated the energy efficiency of buildings and new regulations have been proposed and enforced. The importance of building's energy performance has become even more significant due to the global warming effect and the fact that many countries have signed treaties to reduce their CO_2 emissions.

The early views idealized a building as a closed environment with an indoor climate fully controlled by the most modern HVAC systems. Any type of exchange with the outdoor environment is kept to a minimum to reduce energy loss. Unfortunately this approach led to the emergence of the sick building syndrome. These buildings were sealed against the outdoor environment and therefore the introduction of fresh air was very low. As a

result the concentration of pollutants in the air rose and indoor air quality deteriorated severely.

An increasing attention is now given to the comfort of the building occupants. A building should not just shelter people from the sun and the rain but also provide a comfortable and pleasant working environment. The sealed buildings failed in that. Modern bioclimatic architecture dictates a maximum exploitation of the local climatic and geographic characteristics to minimize energy consumption and provide a comfortable environment. This implies taking advantage of the sun for heating and lighting, natural ventilation for cooling and trees for shading among other things. Mechanical means are only used as supplementary and when no other alternatives are feasible.

Nevertheless bioclimatic architecture is not always an option. There are millions of buildings already constructed and in use. Also, urban construction often poses many restrictions, thus limiting the bioclimatic techniques that can be applied. In either case a need for sophisticated control systems is apparent. The main characteristic of such systems should be the ability to control the indoor environment with whatever means are available, in order to simultaneously achieve two different goals – energy efficiency and occupant comfort.

1.2 Use of controls in modern buildings

1.2.1 Building Energy Management Systems (BEMS)

BEMS is a generic term used to describe computer-based control systems for building services such as air-conditioning, lighting, ventilation, security etc. [1]. The usual layout of a BEMS includes a central station (usually a computer), which provides the BEMS operator with information on the conditions inside the building and the status of the equipment. The operator can remotely respond to problems, program controllers and review the performance of individual equipment or the whole building.

Originally BEMS consisted of a computer hardwired to a number of actuators and sensors. Due to the limited capabilities of computer equipment, BEMS were used solely to provide a central point for system monitoring and basic remote On/Off switching of equipment. With the advancements of technology modern BEMS have also been enhanced. Now the outstations include autonomous intelligent controllers that can be connected to the central stations using a number of communication protocols, even wireless ones. The central station provides, through

graphical, user-friendly interfaces, easy monitoring of equipment and remote programming of controllers.

BEMS provide improved plant performance by achieving the desired operating conditions, by automatic response to failures and by reducing energy consumption through controller fine-tuning. Of course BEMS are associated with a certain design and installation cost as well as a subsequent operating and maintenance cost.

1.2.2 State of the art in building control techniques

There are three main areas of activity in building controller development. Research has been carried out in the use of neural networks, fuzzy systems, predictive control and their combinations [Clarke, et al., 2]. Many of the proposed controllers incorporate provisions for occupant thermal comfort and almost all seek to maximize the building's energy efficiency, either directly or indirectly. Brief reviews of some of the most recent publications on building controllers follow.

The majority of modern controllers use fuzzy logic. [Hamdi and Lachiver, 3] proposed a controller based on human thermal comfort. The controller consists of two fuzzy systems. The first part determines the comfort zone based on current conditions and a user dependent model, while the second one provides the control. An energy conservation of 20%was observed when compared to conventional On/Off control.

[Salgado, et al., 4] tested the performance for heating and cooling of an environmental chamber by fuzzy On/Off and fuzzy PID adaptive controllers. These controllers were able to achieve better trajectory tracking and smaller overshoots when compared to conventional On/Off and PID controllers.

A fuzzy controller was also chosen by [Dounis and Manolakis, 5] for thermal comfort regulation. The controller uses the PMV index and the ambient temperature in order to control heating, cooling and natural ventilation (by means of a window).

[Kolokotsa, et al., 6] developed and tested a family of fuzzy controllers, namely a fuzzy PID, a fuzzy PD and an adaptive fuzzy PD. The controllers were used to regulate thermal and visual comfort as well as air quality inside a building. The inputs used, were the PMV index, the CO_2 concentration and the illuminance level.

On the other hand [Ben-Nakhi and Mahmoud, 7] developed and evaluated a family of six neural networks. They were used to determine the time of the end of thermostat setback in an office building so that by the arrival of the employees the conditions inside were back to normal.

The neurobat project developed by [Morel, et al., 8] uses neural networks for predictive control. The concept behind neurobat is to use neural networks for predicting outdoor temperature and solar radiation. The predicted values are then used by another neural network to forecast building behavior. The resulting, predicted indoor air temperature and a fuzzy estimate of the comfort zone are fed to the controller which operates the heating valves. The neurobat controller was able to reduce energy consumption by 35% in comparison to a standard commercial controller and 11% in comparison to a "performant" commercial controller. At the same time the controller was able to improve thermal comfort.

Neural-fuzzy systems have also been studied. [Egilegor, et al., 9] tested a fuzzy-PI controller with and without neural adaptation. The system was used to control heating and cooling within the PMV comfort zone. In comparison to the On/Off controller, the fuzzy-PI yielded substantially smaller deviations from the optimum. On the other hand neural adaptation did not offer significant improvement.

[Yamada, et al., 10] developed a controller that uses neural networks, fuzzy systems and predictive control. This controller is used to improve energy saving in air conditioning systems. Specifically it predicts outdoor conditions (air temperature and solar radiation) as well as the number of occupants. These predictions are subsequently used to estimate building performance (air temperature, wall temperature and heat load) in order to determine the heat sources, optimal start/stop times, optimal night purge time during summer and minimum outdoor air intake. In addition to that the controller aims at maintaining indoor conditions within a comfort zone, which is determined by the PMV index.

It is noteworthy that even when the controllers aimed solely at achieving thermal comfort, the results also showed reduced energy consumption. Almost all the controllers used the PMV index as thermal comfort measure. Although PMV is very common its applicability in all types of climates and buildings has been questioned. It will be shown in the following chapter that using the newer adaptive comfort standard (where applicable), it is

possible to achieve higher energy conservation, without compromising occupant comfort.

1.3 Thesis goal

This thesis aims at the design of a controller using the advancements in artificial intelligence and machine learning. This controller should be capable of being incorporated into any building with minimum required modifications and fine-tuning. As such the controller should be able to learn from its environment and be able to adapt itself if the environment changes.

1.4 Thesis organization

The following items provide an overview of each chapter of the thesis.

- Chapter 2 refers to the comfort of building occupants. Comfort is influenced by four main factors: thermal conditions, indoor air quality, lighting and noise. Regarding thermal comfort the Fanger model along with the PMV and PPD indexes are presented. Then the validity of these indexes is investigated and the new adaptive comfort standard is introduced.
- Chapter 3 is an introduction to reinforcement learning. After a description of the main terms used in reinforcement learning, a brief account on dynamic programming, Monte Carlo methods and temporal difference methods is given. Then we focus on the temporal difference methods and several learning algorithms are analyzed. Finally the topic of reinforcement learning with function approximation is addressed.
- Chapter 4 describes the application itself. At first the specifics on the design of the two developed controllers (the reinforcement learning fuzzy controller and the reinforcement learning linear controller) are given. A discussion on the possible ways of selecting a reinforcement signal follows. Afterwards the design of the occupant simulator is reviewed; this simulator will be used to get a direct estimate of the occupant thermal comfort.
- Chapter 5 presents the results of the simulator and the controllers for a variety of testing configurations. The controllers are compared to each other and to a fuzzy-PD controller and an On/Off controller.
- A discussion on possible problems and future enhancements is the topic of chapter 6.

2 COMFORT

2.1 Introduction

When we refer to the building occupant's comfort, we usually consider several factors such as thermal comfort, indoor air quality, light and noise level. When a person is in discomfort due to any of the aforementioned factors, then a performance drop is expected. It is therefore desirable to regulate the environmental conditions inside the building, in order to achieve comfort and thus increased performance of the occupants.

The need to achieve comfort has been exacerbated by the emergence of the sick building syndrome or SBS. This syndrome is more common in fully air-conditioned buildings and it is usually identified by complaints about stuffy air and discomfort, loss of concentration, weariness and headaches from the building's occupants. High temperatures, low humidity, noise, insufficient illumination and inadequate ventilation have been identified as reasons for the occurrence of SBS. In short in all studied cases of SBS the

building's occupants where in discomfort, dissatisfied by the conditions inside the building.

In order to help engineers deal with the issue of comfort in modern buildings, several standards have been published while others are still in development. Thermal comfort is addressed in the ISO 7730:1994 standard as well as in ASHRAE 55. Both standards are based in the work of Fanger and the PMV index that will be discussed later. Relative to the ISO 7730 are the ISO 8996:1990 and ISO 9920:1995 that discuss the issues of metabolic heat production and clothing insulation properties respectively. These two parameters are essential for the evaluation of the PMV index.

The ISO 8995:1989 describes the lighting demands of indoor work environments, while the ISO 1996-3:1987 and the ISO 1999:1990 describe noise limits and the impact of noise to human hearing. Currently there is no ISO standard describing indoor air quality requirements for reasons that will be investigated in a later section.

2.2 Thermal Comfort

2.2.1 Introduction

Thermal comfort plays a key role in assessing comfort. Thermal comfort is defined in the ISO 7730 standard as being "That condition of mind which expresses satisfaction with the thermal environment". This definition, although straightforward, does not provide the means to measure thermal comfort in a specified thermal environment. It is evident that the conditions under which the human body is comfortable with the environment should be investigated.

Although the exact mechanisms that assess and regulate body temperature are quite complex and not fully understood we can use a simple model to illustrate the main processes involved. When the body becomes to warm, it tries to increase heat loss by blood vessel dilation and sweat. Equivalently when it gets to cold the blood vessels constrict and shivering occurs, thus increasing heat production and reducing heat loss. The reactions are triggered when the skin temperature drops below 34ºC or the body temperature rises above 37ºC. As the temperature moves away from these limits the reactions are getting stronger. Therefore a condition for a person to feel thermally comfortable is that the combination of body and skin temperature is between the prespecified limits.

The human body also reacts to the changes of body and skin temperature, especially if they are abrupt and large. Therefore a second condition for achieving thermal comfort is to ensure thermal equilibrium, that is the heat produced in the body should be equal to its heat loss, resulting in the following comfort equation.

$$M - W = H + E_c + C_{res} + E_{res} \tag{2.1}$$

M : Metabolic Rate

W : Effective mechanical power (External work)

H : Dry heat loss

E_c : Evaporative heat exchange at the skin during thermal neutrality

C_{res} : Convective respiratory heat exchange

E_{res} : Evaporative respiratory heat exchange

Although equation (2.1) can be quite difficult to solve, it readily shows the parameters on which thermal comfort depends and therefore the physical quantities that need to be measured in order for estimates of comfort to be made.

Metabolic rate refers to the amount of energy released by body metabolism and is a function of current activity. It is logical that people sleeping or involved in undemanding activities produce far less energy than people involved for example in sports. Metabolic rate is usually expressed in Met ($1\text{Met} = 58.15\,{}^{w}\!/_{m^2\text{ of body surface}}$) and there are tables that provide the metabolic rates of various activities. A table of the met values of various representative activities is provided in the appendix.

External work refers to energy transferred to the human body from the environment and is usually considered to be zero.

In order to determine the dry heat loss, the heat loss due to evaporation at the skin and the heat loss due to respiration we need to know the air temperature, the mean radiant temperature, the air velocity, the humidity as well as the kind of clothing used.

Clothing is significant because it acts as insulation, thus reducing heat loss. The unit normally used for measuring the insulation effect of clothing is the Clo unit ($1\text{Clo} = 0.155\,{}^{m^2 \cdot °C}\!/_{W}$). Just like the metabolic rate there are tables that provide the Clo value of individual garments. Adding the Clo values together we can obtain the total Clo value with adequate accuracy. Notice should be taken to include also the Clo values of seats or beds where

applicable since they also contribute in reducing heat loss. The clo values of several garments are provided in a table in the appendix.

The mean radiant temperature is defined as the temperature of an imaginary black enclosure which would result in the same heat loss by radiation from the person as in the actual enclosure. Since the measurement of the mean radiant temperature is difficult and time consuming we usually resort to the use of one of the operative, equivalent or effective temperatures.

Operative temperature is the temperature of an imaginary room where the air velocity and humidity is the same as in the real room but its ambient temperature is the same as its mean radiant temperature. Equivalent and effective temperatures are defined the same way as the operative but the air velocity in the imaginary room is zero for the equivalent temperature and the humidity is 50% for the effective temperature. It must me noted that the equivalent and effective temperatures depend on the person's clothing and activities, while the operative is normally independent of these parameters [11].

Thermal comfort can be redefined now as the condition where a person is thermally neutral, where thermal neutrality is achieved at a temperature equal to the equivalent temperature, which can be calculated from equation (2.1). Although thermal comfort also depends on humidity it is uncommon to attempt the calculation of a comfortable humidity level since when a person is close to a state of thermal comfort the influence of humidity is small.

In real conditions the building user's clothing, activities and preferences differ from each other and there are also localized phenomena to be considered (like draught), thus making the task of achieving thermal comfort for all occupants virtually impossible [Olesen and Parsons, 12]. Fortunately people are able to regulate their thermal comfort by adjusting their clothing to suit the conditions. So the problem of achieving thermal comfort becomes the problem of minimizing the number of dissatisfied people.

2.2.2 The PMV and PPD indexes

The PMV (Predicted Mean Vote) index is a number that "predicts" the mean thermal vote of large volume of people. This vote is given in a seven grade thermal sensation scale ranging from +3 (hot) to -3 (cold) while zero

represents thermal neutrality. The PMV is calculated using the comfort equation described earlier.

It should be noted that ISO 7730 describes specific ranges of conditions were the use of the PMV index is valid. These ranges are shown in Table 2-1. Also care should be taken when using the PMV because it is based on North American and European, healthy adults in sedentary activity and applying it to different groups may produce deviations.

Table 2-1: The validity ranges of the ISO 7730 PMV index

	Min	**Max**
Air temperature	10ºC	30ºC
Mean radiant temperature	10ºC	40ºC
Air velocity	0m/s	1m/s
Metabolic rate	0.8 Met	4 Met
Clothing	0 Clo	2 Clo

The PPD (Predicted Percentage of Dissatisfied) represents the percentage of people that could be dissatisfied in the given environmental conditions which was defined as the people voting +3, +2, -2 or -3. The relation between the PMV and PPD values is given by equation (2.2) [Memarzadeh and Manning, 13].

$$\mathrm{PPD} = 100 - 95\mathrm{e}^{-0.03353\mathrm{PMV}^4 - 0.2179\mathrm{PMV}^2} \tag{2.2}$$

Figure 2-1 shows the relationship between PMV and PPD graphically. It is evident that the PPD never drops below 5% meaning that there will always be a percentage of dissatisfied. It is also noteworthy that we have a region of PMV values close to zero where the percentage of dissatisfied people is quite low.

In several building simulation, comfort estimation and HVAC control studies the need arises for a long-term thermal comfort measure. For this purpose [Olesen and Parsons, 12] proposed the following index:

$$\begin{aligned} \text{warm period:} \quad & \sum \frac{\text{PPD of actual PMV}}{\text{PPD of upper limit PMV}} \\ \text{cold period:} \quad & \sum \frac{\text{PPD of actual PMV}}{\text{PPD of lower limit PMV}} \end{aligned} \tag{2.3}$$

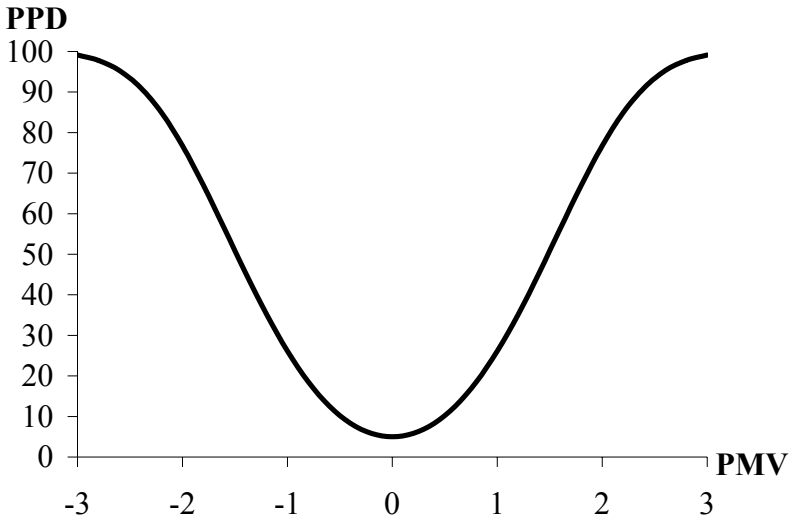


Figure 2-1: Relationship between the PPD and PMV indexes

2.2.3 The validity of the PMV and PPD indexes

Even as back as the 1970s, questions were raised on whether the PMV can accurately predict the AMV (actual mean vote). Although laboratory studies usually supported its validity, field studies often found discrepancies between PMV and AMV. These discrepancies have been attributed from researchers to various reasons.

The Fanger model that is used to derive the formula for the calculation of the PMV is not accurate enough. This model describes the human body as a one dimensional system in steady state. Although more accurate models have been proposed and used in other applications, the Fanger model remains the base of PMV calculations because it is simple enough to be used for thermal comfort assessment of real buildings [Jones, 14].

The use of the metabolic rate and clothing parameters have also been investigated as possible reasons for errors in PMV calculations. Both the metabolic rate and clothing are difficult to measure in field studies and it has been argued that people from different geographical regions may have different metabolic rates for the same kind of activity. On the other hand clothing is determined by using a uniform Clo number that is considered to be equivalent to the effect of the individual garments. Although this may yield reasonable results in several situations, there are considerations as to whether this practice is valid. It must also be noted that the insulation properties of the different garments have been calculated under low air velocity and are thus inaccurate in different conditions.

Psychological reasons have also been suggested. It is plausible that people's expectations may influence their comfort sensation. It has also been demonstrated that when people had a degree of control of their environment, e.g. by operable windows or fans, the comfort range was increased [Dear and Brager, 15]. This may account to the fact that the PMV is quite accurate in buildings with centralized HVAC, but inaccurate in naturally ventilated (NV) buildings [Olesen and Parsons, 12]. These findings led to the development of the Adaptive Comfort Standard (ACS) which will be described in the following section.

Statistical analysis of a large number of data coming from different building across four continents was conducted by [Humphreys and Nicol, 16]. The researchers found that using the data as a whole the PMV showed negligible bias, specifically it was higher than the AMV by 0.11 ± 0.01 scale units. The individual parameters on the other hand showed significant bias in their extremes. It was demonstrated that the PMV overestimated the discomfort of people at temperatures over 27ºC, underestimated the cooling effect of air velocities above 0.2m/s, overestimated discomfort at relative humidity over 60%. They also found out serious bias when the activity is over 1.5 met and when the clothing is outside the 0.3-1.2 clo range. Measurement error was eliminated as a significant factor, since the addition of noise in the measurements had little effect on the biases demonstrated above.

To eliminate the possibility that the individual biases may be neutralized when combined, the researchers conducted tests for each individual building. These tests showed that although the PMV was accurate in moderate environments, it overestimated the discomfort of people in warm and cool environments. This bias became increasingly important when moving to more extreme conditions, regardless of the building type.

2.2.4 Adaptive Comfort Standard (ACS)

The ISO 7730 and ASHRAE 55 standards both acknowledge the fact that people have the ability to adjust their comfort sensation by adjusting their clothing or the local air velocity. Nevertheless both standards failed to take into account the psychological adaptability people exhibit. People living in fully air-conditioned spaces become accustomed to small variations of indoor conditions and may respond negative even in small changes of their environment. On the other hand people living in naturally ventilated buildings are used to large diversity due to daily and seasonal outdoor

climatic conditions [Dear and Brager, 15]. Therefore the latter should exhibit different preferences and wider tolerances.

This psychological adaptation was found to be supported by experimental data of the ASHRAE RP-884 database. In addition to that it was also found that PMV was very accurate at predicting the comfort sensation of people in HVAC buildings, but was very inaccurate in its predictions for naturally ventilated buildings.

In 1978 Humphreys suggested that the optimum internal temperature – the comfort temperature – should be a function of the monthly mean outdoor temperature. Based on the afforementioned inaccuracies of the ASHRAE standard and the suggestion of Humphreys [Dear and Brager, 15] proposed a revision of the ASHRAE 55 standard that accommodates the inaccuracy of the standard for naturally ventilated buildings.

The proposed revision makes use of the adaptive thermal comfort approach to estimate the comfort temperature inside a building. The adaptive thermal comfort model is based on the outdoor air temperature to estimate the indoor comfort temperature. Several researchers have questioned the relevance of these two variables, but it is becoming common agreement that outdoor temperature can influence the behavioral adaptation of people, since the choice of their clothing depends to a certain degree on current environmental conditions. Additionally the weather can influence the expectations of people.

Specifically [Dear and Brager, 15] proposed that the thermal comfort temperature in NV buildings is a function of the outdoor air temperature exclusively. Using the data from the RP-844 database, the following estimation of the comfort temperature was produced.

$$T_C = 0.31 \cdot T_{air,out} + 17.8 \tag{2.4}$$

The comfort zones for 90% and 80% acceptability were also determined to be two bands of 5ºC and 7ºC width respectively.

In a similar study by [Nicol and Humphreys, 17] it was determined that the following comfort equation is a very close approximation to the real one for free-running buildings.

$$T_c = 0.54 \cdot T_{month\,mean} + 13.5 \tag{2.5}$$

Regarding the comfort zones, it is argued that their width depends on the availability of control for the building occupants.

In 1997 the EU funded a program, whose objective was to provide a method of reducing energy consumption in buildings, by utilizing the ACS. The program lasted 3 years with participants from UK, France, Greece, Sweden and Portugal. During the project thermal comfort studies were carried out in various buildings in all the participating countries. The data gathered were used to determine a comfortable temperature as a function of outdoor temperature [McCartney and Nicol, 18]. Instead of the monthly mean, a running mean temperature was preferred. The running mean was calculated by:

$$T_{RM,n} = cT_{RM,n-1} + (1-c)T_{DM,n-1}$$
$$T_{RM,n} \text{ is the running mean of day n} \qquad (2.6)$$
$$T_{DM,n} \text{ is the daily mean of day n}$$

The c constant was chosen to be 0.80 after testing. Using regression analysis [McCartney and Nicol, 18] obtained the following thermal comfort estimate:

$$\begin{aligned} T_C &= 0.302 \cdot T_{RM} + 19.39, & T_{RM} &> 10^{\circ}C \\ T_C &= 22.88, & T_{RM} &\le 10^{\circ}C \end{aligned} \qquad (2.7)$$

The comfort equations as calculated for each country are presented in Table 2-2.

Table 2-2: ACS comfortable temperatures for 5 EU countries

Country	$T_{RM} \le 10^{\circ}C$	$T_{RM} > 10^{\circ}C$
France	$T_C = 0.049 \cdot T_{RM} + 22.58$	$T_C = 0.206 \cdot T_{RM} + 21.42$
Greece	N/A	$T_C = 0.205 \cdot T_{RM} + 21.69$
Portugal	N/A	$T_C = 0.381 \cdot T_{RM} + 18.12$
Sweden	$T_C = 0.051 \cdot T_{RM} + 22.83$	N/A
UK	$T_C = 0.104 \cdot T_{RM} + 22.58$	$T_C = 0.168 \cdot T_{RM} + 21.63$

Although the comfortable temperatures differ from study to study the ACS is very important. The correlation between the results given by the ACS and the AMV is better than that between PMV and AMV in NV buildings. Also it is noteworthy, that in several applications of the ACS in real buildings significant energy conservation was observed, without adverse effects on the occupant's thermal comfort.

2.3 Indoor Air Quality (IAQ)

2.3.1 Introduction

Although air pollution has been a topic of research and development for almost a century now, the issue of IAQ is quite recent. Indoor spaces, up until recently, were considered safe and free of pollutants. It is characteristic of this misconception that during the 1970s people were encouraged to build well insulated buildings to conserve energy and to keep the harmful, polluted air outside.

Today we know that indoor air concentrations of various irritating, carcinogenic and mutagenic compounds are larger than their corresponding outdoor concentrations even in industrial areas.

Since people spent 70% to 90% of their life indoors, indoor air quality poses a significant threat to health. Several organizations including the European Union, the U.S. Environmental Protection Agency and the World Health Organization have conducted studies to catalog the pollutants and their effects.

Indoor air pollutants may originate from the outdoor environment, building materials, an indoor activity like painting or smoking, even the earth. The problem is that these pollutants disperse in a relatively small space, from which it is difficult to escape, thus they maintain high concentrations. These concentrations may increase with the increased temperature and humidity, often encountered in indoor spaces. In addition to this the presence of a large number of different compounds, possibly interacting, in indoor atmosphere can have unpredictable effect on human health.

2.3.2 Common indoor pollutants

The variety of indoor air pollutants is large and their presence depends on the geographical region where the building is situated, the building materials, on the type of activities practiced and the environmental systems (air conditioners, air purifiers, window types etc.) used. Most common pollutants are radon, cigarette smoke, volatile organic compounds and biological products.

Radon is a carcinogenic compound that usually originates from the earth or the building materials. It is a very important factor contributing to lung cancer, second only to smoking. Although special means to overcome the problem, like insulating materials, are available, the cost can only be

justified in the case of known high concentrations. In most cases good ventilation is the only available solution.

Cigarette smoke is a very common indoor air pollutant. It contains 4000 chemical compounds, more than 40 of which are known carcinogens. Cigarette smoke can be divided into two streams, the first inhaled by the smoker and the second that is dispersed in the environment. Exposure to this second stream is known as passive smoking. Its composition depends on the number of people smoking and the kind of cigarettes they smoke. Passive smoking is proven to be more harmful than smoking since the passive smoker inhales 1 to 50 times the quantity of various carcinogens compared to the smoker. It is noteworthy that CO concentrations may reach 50ppm in a room with many smokers, while the normal outdoor concentration in the centre of Athens is 5ppm.

A very common class of indoor pollutants is that of biological origin, from plants, animals or even humans. They are very common and usually emerge wherever the humidity is high. Several diseases, some of which possibly lethal like legionnaire's disease, are related to these kinds of pollutants. A good practice to keep these pollutants under control is to keep indoor humidity between 40% and 60%. Rooms with high humidity like bathrooms and kitchens should be ventilated often and all surfaces should be kept as dry as possible.

Combustion products like CO_2, CO, SO_x and NO_x are also found in indoor air especially when badly regulated devices (stoves, fireplaces etc.) are in operation. High concentrations of any of these pollutants is dangerous so all fuel burning devices should be carefully maintained to eliminate their indoor emissions. Fortunately the number of such devices is very small and they are usually found only in residential areas.

Several products of daily use, like detergents, disinfectants, paints, cosmetics, etc. contain volatile organic compounds (VOC), which during use or storage are introduced into the indoor air. More than 300 such compounds have been catalogued, several of which are known carcinogens, neurotoxins or respiratory irritants. The problem can be alleviated with good ventilation especially during activities such as painting, where the concentrations of these compounds may reach dangerous levels.

Ozone is another dangerous pollutant of indoor air found usually in office spaces where copying machines or air cleaners are in use. Fortunately

ozone concentrations rarely reach higher values than that of outdoor urban air.

Several other pollutants may also be found in indoor air like asbestos, formaldehyde and lead. Some of these are found only in special circumstances (e.g. asbestos is found in old buildings with asbestos insulation) while others are quite common.

2.3.3 Controlling indoor air quality

It has already been mentioned that a large number of possibly interacting pollutants are found in indoor air. Although for several of them studies have been carried out as to their effects and safe concentration limits, these studies usually apply to industrial areas and are difficult to incorporate them into building regulations and standards concerning the commercial and residential sector.

It is unfeasible to install measuring devices that can keep track of the concentrations of all these compounds and even when these devices are available it is very difficult to keep their levels under control. The means that are available to improve air quality are the use of air cleaners and ventilation. The former is of limited use, since air cleaners cannot handle all types of pollutants and are quite costly. The latter has been proven to be a better solution and many regulations and standards resort to defining minimum ventilation rates in order to maintain IAQ. Of course ventilation is also associated with a certain cost since fresh air is usually in a different temperature than the one desired.

In order to assess IAQ the CO_2 concentration is often used. Although CO_2 does not account for the sum of all the indoor air pollutants, its concentration can be used as a measure of ventilation efficiency. For example if the CO_2 concentration rises above the 800ppm threshold, a IAQ-sensitive controller would be instructed to increase ventilation or to increase the fresh air intake ratio in a central HVAC system.

2.4 Light requirements

2.4.1 Introduction

Light plays a significant role on the overall human comfort and on their performance. In order to achieve light requirements inside a building, besides artificial light, also the natural light from outside is used. The use of natural light, known as daylighting, has been a topic of architectural design since man first built dwellings. Daylighting has received greater

attention the last 40 years because, besides the aesthetic advantages it provides, it is also a good way to conserve energy. Some control over daylighting is available, for example by operable shading devices, but it is more common to control only artificial light.

2.4.2 Terms

In lighting four units are in use:

- *Luminus flux (Φ).* Refers to the amount of light per unit of time. Measured in lumen (lm).
- *Luminus intensity (I).* Measures flux in a given direction. The unit used is the candela (cd).
- *Luminance (L).* Indicates the amount of lightness of an emitting surface and is measured in cd/m^2.
- *Illuminance (E).* Refers to the flux reaching a given surface and is measured in lux (lx).

2.4.3 Visual comfort

There are three factors that affect visual comfort, illuminance, glare and light color as analyzed in [Serra, 19]. Illuminance refers to the adequacy of light to perceive the objects of out interest. Glare is an unpleasant effect that occurs when points of excessively different luminance are present in our visual field. This may result in the generation of optical artifacts (e.g. rays or coronas around very bright objects in dark backgrounds), inability to distinguish some objects whose luminance is very different from the mean, even inability to see (incapacitating glare) when a beam of light strikes the center of the eye. The color of the light is also an important factor of comfort. Color is measured by a temperature. This temperature corresponds to the temperature of black body that emits the same light. Temperatures below 5000K are reddish and considered warm while those above 5000K are blue and considered cool. The temperature of natural light is 6000-6500K and is considered very good from the optical comfort standpoint. Nevertheless natural light can also cause discomfort in the case of low light levels since it is too cool [Serra, 19].

The regions of comfort for each of these parameters depend on the activity types of the individuals. It is natural that the light requirements are different when an activity requires precision and concentration and when a person is just moving around a room. In the appendix, tables are provided that give indicatory ranges for some types of activities.

Light requirements are usually assessed during the design of a building so that the artificial lighting system provides adequate light of the appropriate temperature for the proposed use of the building. Control systems are usually employed to adjust artificial lighting as required and in some cases to regulate shading devices (e.g. curtains) in order to avoid glare.

2.5 Noise

Noise is also an issue when studying comfort since it is known that people in noisy environments have concentration problems and increased stress. It is also a fact that people exposed in noisy environments for long periods of time will experience hearing loss in the future.

Achieving sound level requirements is an issue of building design rather than a control problem and therefore only a brief description of the noise requirements is presented in this thesis.

Both ISO and ASHRAE have developed standards describing the permissible noise exposure. For example Table 2-3 shows the permissible noise exposures as described in the Occupational Safety and Health Standard 1910.95 of the US Department of Labor.

Table 2-3: Permissable noise exposures (Occupational Safety and Health Standard 1910.95 – US Department of Labor)

Duration (hours per day)	Sound level (dBA)
8	90
6	92
4	95
3	97
2	100
1.5	102
1	105
0.5	110
0.25 or less	115

Of course the standards refer mostly to industrial workspaces and workspaces with high noise levels. In office and residential buildings the noise is usually much lower. Sound levels inside buildings are usually measured in terms of Noise Criteria (NC) values. The higher the NC number, the higher the sound level. ASHRAE suggests a NC 35 – 40 level for offices [Carrier, 20].

In some situations it is even desirable to have noise. For example many open offices have background sound masking, where a “white noise” is introduced to mask conversations between cubicles, so that the employees are not distracted by each other. These systems are often set to an NC value of 41 – 43.

3 REINFORCEMENT LEARNING

3.1 Introduction

Reinforcement learning refers to a variety of learning algorithms that are suitable to solve the problem of learning by evaluating the response of the environment to the actions taken by the learning agent. In contrast with the supervised learning problem where the learning agent is told what the best action was for any given situation, in the reinforcement learning problem the agent is only given a numerical reward showing how "good" or "bad" its actions were. Learning occurs only using past experience. The agent must try different actions in order to determine which are good – which maximize its reward. In reinforcement learning problems it is also possible that an action may influence not only immediate rewards but also one or more of the subsequent rewards. For example in a board game a "bad" move during the first stages of the game may severely decrease the chances for winning.

Reinforcement learning is based on the works on artificial intelligence going back to the beginning of the 20th century. These early works, that involved psychology and animal learning, introduced such ideas as trial-and-error and the tendency to selecting actions based on which produced greater satisfaction in the past. During the mid of the 20th century work on reinforcement learning was sparse as the majority of the research involved supervised learning problems. Reinforcement learning became a distinct active research field only during the 1980s mainly by the efforts of Richard Sutton and Andrew Barto which were followed and/or complimented by Tesauro, Bertsekas, Tsitsiklis, Werbos, Watkins and many others. One of the most complete introductory texts on reinforcement learning is [Sutton and Barto, 21].

3.2 Reinforcement learning basics

3.2.1 Terms

At first the main terms used in reinforcement learning will be discussed; these terms are actions, states, policy, reward function, value function, return, discounting, episodic and continual tasks, backup and the Markov property.

Actions refer to the decision that the agent will be called to make and can be as low-level as the voltage applied to a motor unit or as high-level as where the agent should focus its attention on. State on the other hand refers to the available information that is pertinent to the agent's decision making. State can be comprised of any kind of information ranging from sensor signals to symbolic characteristics of the environment.

Policy defines the way a reinforcement learning agent behaves. It provides a mapping between the situations the agent can find itself in (states) and the action it should take. Policies can be deterministic by specifying which action should be taken under each state or stochastic when for example instead of a specific action, probabilities of choosing several actions are given. Judging from the above, the reinforcement learning problem becomes the problem of determining the optimal policy, the policy that will collect the maximum reward in the long run.

The reward function describes the expected reward of being in a certain state or choosing a certain action while being in a specific state. The reward function can be said to be "short-sighted" as it looks only one step ahead. In contrast to the reward function, the value function defines the total amount of reward that the agent should expect to receive in the long-term by being

in a specific state or by choosing an action while being in a specific state. Value functions are very important in determining the optimal policy. Specifically when an exact model of the environment is available the agent can determine which action will result in the best successor state. The best successor state is defined as that with the largest value. Alternatively in problems were a precise model of the environment is not available, the state-action value is used instead since it provides the means to selecting the actions.

A return is the actual reward received by an agent while following a certain policy. The return may refer to the total reward received or to the reward received after a small amount of time. The return can be used to update the value function because it is in fact an estimate of the value function taken from interaction with the environment.

In many situations the reinforcement learning problem may continue indefinitely and the return may reach infinity. In order to overcome this problem and ensure the boundedness of the return, discounting is used. Discounting assigns greater weight to immediate rewards and less to very distant ones. The discounted return can be written as:

$$R_t^{\lambda} = r_t + \gamma r_{t+1} + ... + \gamma^k r_{t+k} + ... \quad (3.1)$$

The γ parameter is known as the discount factor and takes values in the [0 1] interval. The smaller the value the less we care about long term rewards.

All reinforcement learning problems can be divided into two categories: episodic and continual. Episodic problems are problems that have one or more terminal states. When the agent reaches one of these states the episode finishes and the state is reset to its initial setting. An example of an episodic problem is a game of chess, where each episode refers to a single game and the terminal states refer to board positions where either player has won or the game is tied. Continual problems, on the contrary, never terminate but continue indefinitely. An example of a continual problem is a process control problem.

Backups refer to the way value function updating occurs. Specifically it refers to which values are used for updating the current value function. For example when using a one-step backup, the agent looks only one step ahead, that is it uses the value function of only the next state (or state-action) to update the value of the current state (or state-action)

The Markov property is an important property regarding the state signal, since the applicability and performance of many reinforcement learning algorithms depends on it. In order for a state signal to have the Markov property, the environment dynamics must depend only on the current state and chosen action, thus enabling us to predict the next state and its expected reward only using currently available information and not the entire history up to the current situation. Even when a state signal does not have the Markov property, it is desirable that it represents a good approximation of a Markov signal, because in a different case the reinforcement learning system's performance will be poor.

A reinforcement learning task that satisfies the Markov property is called a Markov Decision Process (MDP). For finite MDPs the probability of the occurrence of a possible successor state s´ given a state s and an action a is given by:

$$P_{ss'}^{a} = \Pr\left\{s_{t+1} = s' \mid s_t = s, a_t = a\right\} \tag{3.2}$$

and the expected reward by:

$$R_{ss'}^{a} = E\left\{r_{t+1} \mid s_t = s, a_t = a, s_{t+1} = s'\right\} \tag{3.3}$$

3.2.2 Balancing exploration and exploitation

One of the main characteristics of reinforcement learning is the tradeoff between exploration and exploitation. At any given time the agent must decide whether to choose the best action based on its knowledge or to try something else in order to make better selections in the long-term. In any case the agent has to try all actions several times in order to evaluate their performance in producing large rewards.

In order to balance exploration and exploitation the ε-greedy algorithm is frequently used. Although there are several variations, ε-greedy action selection can be described as always selecting the best action, except in a small fraction ε of the time where other (suboptimal) actions are tried. The variations of ε-greedy algorithms include cases where the ε is not constant but changes with time or where the suboptimal action selection is not random but may depend on the expected reward of an action (e.g. by assigning greater probabilities to better actions) or on the certainty of the estimate by choosing actions that have not been adequately explored.

3.3 Reinforcement learning methods

There are three major categories of reinforcement learning methods, each with its application scope, advantages and disadvantages. There are Dynamic Programming (DP) methods, Monte Carlo (MC) methods and Temporal Difference (TD) methods.

3.3.1 Dynamic Programming

Dynamic programming is a family of algorithms that aim to determine optimal policies given a perfect model of the environment as a MDP. In the majority of DP applications, it is assumed that the state and action spaces are discrete or that they have been quantized.

The main concept behind DP is the search for the optimal policy using estimates of the value functions. It has already been mentioned that an optimal value function can be directly used to determine an optimal policy. In order to find the optimal value function DP uses the Bellman optimality equations as an update rule.

$$V(s) \leftarrow \max_a \sum_{s'} P^a_{ss'} \left[R^a_{ss'} + \gamma V(s') \right] \tag{3.4}$$

The problem behind DP is the need for an exact model of the environment, since each update requires a full backup, that is an evaluation of all possible successor states regardless of which will actually appear. DP methods are also considered impractical for problems with large state spaces since in such a situation a single backup will have large computational requirements.

3.3.2 Monte Carlo

In contrast to DP, MC methods do not require a model of the environment. Instead learning occurs by taking advantage of sample sequences of state, actions and rewards from real or simulated interaction of the agent with its environment.

MC methods in reinforcement learning operate by averaging the returns of a policy in order to estimate the value functions. Depending on which returns are averaged, MC methods are divided into two categories. First-visit MC averages the returns that followed the agent's first visit to state s, in order to determine its value function V(s). On the other hand every-visit MC averages the returns that followed after every visit to s. The same applies if we use state-action values instead of state values. Given the new value function estimates the agent can improve its policy.

The value updating rule is of the form:

$$V(s_t) \leftarrow V(s_t) + \alpha\left[R_t - V(s_t)\right]$$
$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots \tag{3.5}$$

An important advantage of MC methods is that the value estimates for each state are independent and therefore the computational demands of a acquiring a value estimate of a single state is small.

MC methods suffer from two major drawbacks. The first is that MC methods can only be applied to episodic tasks since learning takes place only after the end of an episode and not incrementally. The second is that when a deterministic policy is in use (e.g. greedy), there is a good chance that several actions will not be tried and therefore their corresponding value estimates will not be improved. It is therefore necessary to ensure sufficient exploration.

3.3.3 Temporal Difference

Temporal Difference learning methods combine the advantages of MC and DP methods. They do not require a model of the environment and they are able to learn from interaction with the environment on a step by step basis.

In order to make an update of the state value function, TD methods only require the observed reward and an estimate of the value of the next state. The update rules used in TD learning are of the following form:

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right] \tag{3.6}$$

Using an estimate $\left(V(s_{t+1})\right)$ to improve another estimate $\left(V(s_t)\right)$ is called bootstrapping and is a major characteristic of both TD and DP methods. Bootstrapping gives TD methods the capability of online implementation, by taking advantage of whatever knowledge is available as soon as it is available.

TD methods, under certain assumptions, have been proven to converge to an optimal policy and it is also true that in several applications they have been found to converge faster than MC methods [Sutton and Barto, 21].

To sum up, we have already seen that both MC and TD methods use value update rules of the form:

$$\text{new\_value} = \text{old\_value} + \alpha\delta \tag{3.7}$$

where δ is the value error. The MC error is defined as the difference between the current value estimate and a full return, while the TD error is

the difference between the current value estimate and a discounted return after observing the next state.

3.4 Temporal credit assignment in TD

In the previous section we discussed that TD methods learn on a step by step basis using a one-step backup. On the other hand MC methods use a multi-step backup up to the terminal state. It is possible to use TD algorithms that perform multi-step backup. This is expected to increase the speed of the algorithm since from a single experience several states, visited in the past, will be updated. In order to achieve multi-step backups online, eligibility traces are used.

Eligibility traces can be seen from two viewpoints, the forward and the backward [Sutton and Barto, 21]. The forward view is more theoretically oriented and it states that eligibility traces represent how far ahead and which states should we look in order to determine the current best action. The backward view is oriented towards implementation and it states that eligibility traces represent a memory of which state (or state-action) values are "eligible" for updating due to the currently received reward.

3.4.1 n-step TD

There are various ways to implement eligibility traces. Perhaps the most simple is the n-step TD. An n-step backup is based on the first n rewards and the estimate state value n-steps ahead. For example a 3-step TD update rule would be of the following form:

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V(s_{t+3}) - V(s_t)\right] \qquad (3.8)$$

Examining equation (3.8) we can determine that the term $r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V(s_{t+3})$ is in fact an approximation of the full reward used by the MC methods. Actually the larger the n value, the closer we get to accurately approximating the full backup of MC methods.

Despite their simplicity, n-step TD are seldom used due to their implementation problems especially for large n values, where the agent must wait a long time before an update is made.

3.4.2 Complex backups

Complex backups refer to backups that average in any way two or more n-step backups. For example a backup can be done towards the average of a two step return and a four step return. In order to facilitate implementation of these complex backups the TD(λ) algorithm was developed. The λ

constant is a parameter that defines the weighting of each backup. The return used is defined by:

$$R_t^{\lambda} = (1-\lambda)\sum_{n=1}^{\infty} \lambda^{n-1} R_t^n \tag{3.9}$$

From equation (3.9) we determine that TD(0) is the simple TD method we discussed in the previous section and TD(1) refers to a variant of the MC algorithm.

3.4.3 Implementing TD(λ)

In order to implement the TD(λ) algorithm we need to keep track of the eligibility trace of each state e(s), in addition to its value. The update rule of TD(λ) is just like that of TD(0) but for the dependence on the eligibility trace. It should also be noted that now we update all states and the eligibility trace takes care of not updating irrelevant state values.

$$V(s_t) \leftarrow V(s_t) + \alpha\left[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\right] e(s_t) \tag{3.10}$$

Initially the eligibility traces of all states are zero. Every time step the eligibility traces are updated according to the following rule:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & s \neq s_t \\ \gamma\lambda e_{t-1}(s) + 1 & s = s_t \end{cases} \tag{3.11}$$

The former update rule is known as the update rule for “accumulating eligibility traces”. Also of use are the “replacing eligibility traces” whose update rule is:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & s \neq s_t \\ 1 & s = s_t \end{cases} \tag{3.12}$$

The difference between accumulating and replacing traces is that in the former repeated visits to a state will increase its eligibility trace thus producing a great change. Although this may seem prudent, in practice it has shown to result in poor performance and in some situations it may cause serious error [Reynolds, 22]. For example in a case where an agent makes repeatedly wrong actions and at some point takes the correct action and ends up in a terminal state with large reward, the wrong actions’ values will be corrected towards the reward more than the correct action.

It can be shown that when we apply the TD(1) algorithm, the use of accumulating traces gives the every-visit MC while replacing traces gives the first-visit MC [Reynolds, 22].

3.5 TD Learning algorithms

In the former sections we discussed various methods of reinforcement learning, we will now focus on some learning algorithms based on TD since it is more suitable for the aims of this thesis. Although up until now we used the state value function V(s) from now forth we will use the state-action value function Q(s,a). This value function is more suitable for control applications were no model of the environment is available. Fortunately all update rules apply to Q(s,a) with no change.

Care should be taken though to the way the eligibility traces are updated, because now the actions should also be taken into account. One possibility is to use the following schema:

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & s \neq s_t \\ \gamma\lambda e_{t-1}(s) + 1 & s = s_t \quad a = a_t \\ 0 & s = s_t \quad a \neq a_t \end{cases} \tag{3.13}$$

3.5.1 Short-sighted

The short-sighted algorithm is not really used since it refers to an algorithm that does not seek to maximize long-term rewards but only the next expected reward. The value function estimate is therefore updated by using just the reward received by taking an action:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} - Q(s_t, a_t)\right] \tag{3.14}$$

3.5.2 Sarsa

One-step Sarsa operates by choosing an action a by applying a policy like ε-greedy to the state-action value. The outcome of that action is then used to update the state-action value estimate according to the following update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right] \tag{3.15}$$

Sarsa with eligibility traces – Sarsa(λ) – uses the same update rule, only now the update also depends on the eligibility of the state-action value.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)\right]e(t) \tag{3.16}$$

It is evident from the update rule that the value updates of Sarsa depend on the policy being followed since the value depends on the next action taken.

3.5.3 Q learning

One-step Q learning operates in much the some way with Sarsa only now the update of the value estimate does not depend on the policy being followed but on the greedy policy. The update rule for one-step Q learning is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right] \quad (3.17)$$

Q learning with eligibility traces is more difficult to implement because when an agent is not following the greedy policy with respect to Q(s,a) then the $\max_{a_{t+1}} Q(s_{t+1},a_{t+1})$ values will not be available. There are two variants of the Q learning with eligibility traces Watkin's Q(λ) and Peng's Q(λ). Watkin's Q(λ) works the same way as Sarsa(λ) but sets the eligibility traces of all states to zero when an exploratory action is taken. The update rule is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_{a_{t+1}} Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)\right]e(t) \quad (3.18)$$

Peng's Q(λ) aims to remedy the problem of losing the eligibility traces when an exploratory action is taken and it is a mixture of Sarsa(λ) and Q learning. In effect all the rewards are updated according to the Sarsa(λ) rule, with the exception of the last which is updated according to the Q learning rule. Peng's algorithm is more difficult to implement but studies have shown that it performs better than Watkin's algorithm.

3.6 Reinforcement learning and function approximation

The reinforcement learning techniques reviewed so far worked by keeping a record of the value estimates of states (or state-actions). Although straightforward, this approach will lead to large memory requirements when the state and/or action spaces are large, in other words these techniques suffer from the curse of dimensionality. Besides that there is no direct way of applying them in continuous space, which is the common case in real applications.

In order to overcome the formerly mentioned limitations, a reinforcement learning agent should be able to generalize, that is to assign similar value to similar situations, even when it has never experienced them exactly. This can be achieved using a variety of approximation techniques.

By reviewing the update rules of the various algorithms it is obvious that, during each update the value estimates where moved towards a discounted return like $R_t = r_{t+1} + \gamma V(s_{t+1})$. This update can be interpreted as an example of the desired input-output behavior of the agent. This means

that the reinforcement learning problem becomes a supervised learning problem. If we consider that the state value function is a parameterized function with parameter vector $\vec{\theta}$, by adjusting the parameter vector, it is possible to find the one, which best approximates the state value function.

Unfortunately not all function approximators are suitable for application in reinforcement learning problems. An essential characteristic of such an approximator would be the ability for online incremental learning, since in the majority of the reinforcement learning problems the training data originate from the agent-environment interaction. In several applications it is also desirable that the agent can function in non stationary environments. Besides that approximators that may work very well in one problem, may fail to converge in other or even in the same task given different parameters.

A general gradient-descent method for the state value prediction problem is given by equation (3.19):

$$\vec{\theta}_{t+1} = \vec{\theta}_t + a\left(R_t - V_t(s_t)\right)\nabla_{\vec{\theta}_t} V_t(s_t) \tag{3.19}$$

The above equation, given enough training samples, will minimize the mean square error (MSE) of the $R_t - V_t(s_t)$ term. Although it can be argued that the MSE may not be an appropriate measure since what we really seek is to find the best policy, there are no other readily available, better alternatives for use [Sutton and Barto, 21]. Given that R_t is an unbiased estimate of the state value function, as is true in the case of the MC return, then the gradient descent method described above converges to a local optimal.

3.6.1 TD learning with function approximation

Equation (3.19) can be used for TD learning where the R_t is the bootstrapped estimate of the value function. Unfortunately since the bootstrapped estimate is not an unbiased estimate there is no guarantied convergence. The update rule of the gradient descent TD(λ) is given by the following equation:

$$\begin{aligned} \vec{\theta}_{t+1} &= \vec{\theta}_t + \alpha\delta_t e_t \\ \delta_t &= r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \\ e_t &= \gamma\lambda e_{t-1} + \nabla_{\vec{\theta}_t} V_t(s_t) \end{aligned} \tag{3.20}$$

3.6.2 Linear function approximation

One of the most popular category of approximators is the linear one. Linear approximators offer the advantages of simplicity, well understood training algorithms and convergence to the global optimum if convergence to a local optimum can be guarantied.

A linear approximator of the state value V(s) can be described by the equation (3.21):

$$V_s(s) = \vec{\theta}_t \vec{\phi}_s = \sum_{i=1}^{n} \theta_t(i)\phi_s(i) \tag{3.21}$$

where $\vec{\phi}_s$ is a vector of features that depends on the state. Consequently the gradient of the value function in equation (3.20) is $\vec{\phi}_s$.

Therefore in order for linear TD(λ) to be employed, there is only the need of determining a method of feature construction. There are several such methods that convert the input to binary or real valued features like coarse coding, tile coding, radial basis functions and others.

The linear approximator described above can also be used directly for control problems if instead of the value function we use the action-value. Special care should be taken though, so that the feature vector contains such vectors that not only depend on the actions but can also capture the relationship between states and actions [Sutton and Barto, 21].

4 APPLICATION

4.1 Introduction

The objective of this application is to design an adaptive controller that will take into account user preferences, in order to achieve energy conservation and user comfort. User comfort entails three distinct goals, thermal comfort, indoor air quality and adequate illuminance. In order to simplify controller design the latter goal will be separated from the other two.

In order to put as few restrictions to the application of the controller as possible, we will assume that an accurate model of the environment is not available. The only information that is available to the controllers is simple sensor measurements like indoor temperature, humidity and others in addition to a user response to the current environmental conditions inside the building.

In order to assess user comfort the controller should be capable of using any of the comfort measures described in an earlier chapter, like the PMV

index or the ACS. In addition to that the controller developed will also be able to use direct feedback from the user. Building users are unlikely to frequently report their comfort status. In addition to that we can not expect them to report feeling comfortable, even if they are given the power to do so. Therefore user response should be modeled by an irregular discomfort signal.

4.2 The Reinforcement Learning Fuzzy Controller (RLFC)

The use of conventional fuzzy controllers although simple and computationally undemanding presents a serious problem regarding the fuzzy rule base generation. Usually the rule base is designed using expert knowledge and trial-and-error techniques which are time consuming and do not provide an update scheme when the environment changes. Reinforcement learning controllers on the other hand choose actions based on a reward function or look-up table. This reward is based on the controller's previous experience.

The design concept of the RLFC is quite simple. Instead of using a fuzzy rule base, a reinforcement learning algorithm is used to associate inputs or states to proper outputs or actions by means of a reward table. The difference between the RLFC and the conventional reinforcement learning paradigm is that the current conditions are not described by a single state, but by several "active" states with varying "activity levels". Choosing an action is then the task of finding the best actions for each active state and aggregating them based on the corresponding activity levels. This methodology allows experience to be passed to several state-action pairs.

4.2.1 Controller operation

The FRLC was developed for Simulink using the Matlab C language and operates in several steps as depicted by the flowchart in the appendix. During initialization the controller reads a binary controller file that contains the fuzzy structure and the rewards.

The controller operates in a different frequency than the rest of the simulation namely every 10 (or possibly more) minutes. Since conditions inside a building are usually slow changing, a 10 minute delay between reevaluating the controller response will not result in inefficiency but on the contrary will help avoid frequent switching on and off of the equipment.

Each time step the controller goes through four separable phases: state-action search, ε-greedy action selection, defuzzification and value updating.

The input to the controller consists of the current state vector, the learning rate, the penalty and an ε parameter that controls the ε-greedy algorithm. Finally after the termination of the simulation the binary controller file is updated using the new rewards.

4.2.2 State-action search

In order to determine which states are active the controller iterates over all possible states. For each state component a membership value is determined and the minimum over all state components is used as the state's activity level. If this activity level is greater than a threshold, then the state is considered to be active. The use of a threshold greater than zero is dictated by the use of gaussian bells as membership functions, that have always non zero value.

When an active state is found the controller iteratively searches all possible actions to find the one with the highest value – expected reward. The state-action combination, along with its activity level is appended in a list. This list is used to update the eligibility trace matrix using accumulating traces:

$$e(s,a) \leftarrow e(s,a) + m(s) \tag{4.1}$$

The eligibility trace update differs from the classic accumulating traces update used in reinforcement learning in that it increases the eligibility traces of all active states by their activity level instead of a single state by one.

4.2.3 ε-greedy action selection

Since the controller's environment may change (e.g. a new AC unit is installed or the occupants change) it is desirable that the controller will continue to try suboptimal actions in order to be able to respond correctly in the event of a change in the environment. This is achieved using ε-greedy action selection.

There are two possibilities for the implementation of the ε-greedy algorithm. The first is to apply it after the defuzzification, on the final selected action. Special care must be taken though, so that the new action's expected reward is high enough and inappropriate actions, like turning on cooling during a cold winter day, are not chosen.

The second possibility, and the one actually used, is to apply the ε-greedy algorithm to the best action for each active state. Every item in the active state-action list is considered for alteration with a probability ε. For

each eligible for alteration state-action pair a randomly chosen action component is changed to a new also random value. Since the final action will always be influenced towards the optimal by other active states with unchanged actions, there is no need to keep a watch for inappropriate actions.

4.2.4 Defuzzification

The controller was implemented with the ability to perform two kinds of defuzzification, namely center of area (COA) and mean of max (MOM). Although COA is probably the most popular method, the most suitable for the current application is MOM. This is because usually the control variables cannot take real values but only one from a collection of settings. MOM defuzzification operates by, in effect, evaluating the state-action list and choosing the setting that is most voted for, the vote being weighted by the activity level of each state.

4.2.5 Value update

During this phase the controller updates the value matrix if each state-action combination. The value update rule is analogous to the TD rules discussed in the previous chapter:

$$\begin{aligned} &\text{new\_value} = \text{old\_value} + \alpha \cdot \delta \cdot e \\ &\alpha: \quad \text{Learning rate} \\ &\delta: \quad \text{TD error (Depends on the learning algorithm)} \\ &e: \quad \text{Eligibility trace (In RLFC the eligibility trace also depends on the activity level of each state visited)} \end{aligned} \tag{4.2}$$

Specifically the TD error is calculated using the following equations:

$$\begin{aligned} &\text{SARSA}: \quad TDerror = r_t + \gamma \frac{\sum m(s')Q(s',a')}{\sum m(s')} - \frac{\sum m(s)Q(s,a)}{\sum m(s)} \\ &Q-\text{learning}: \quad TDerror = r_t + \gamma \max_{a'} Q(s',a') - \frac{\sum m(s)Q(s,a)}{\sum m(s)} \end{aligned} \tag{4.3}$$

where m(s) is the membership – activity level of the state s.

After the value matrix is updated the corresponding eligibility trace matrix is also updated using the following equation:

$$e(s,a) \leftarrow \gamma\lambda e(s,a) \tag{4.4}$$

The update rule is the classic update rule used frequently in reinforcement learning applications with only one difference. The update depends on the activity level of the current state-action, thus the value

change depends also on the influence of each state-action on the final action selection.

4.3 Controller design

Since the value matrix contains all possible combinations of states and actions, a large number of inputs and outputs and/or a fine partitioning of the input space will result in a very large value matrix. For example using the structure of a simple fuzzy controller that functioned efficiently in this problem, resulted in a value matrix of more than $3 \cdot 10^8$ values. A matrix of that size is difficult to handle and very slow to learn. On the other hand if the partitioning of the input space is very coarse the controller will be unable to identify different conditions, thus behaving inadequately. Therefore special care should be given during the controller design phase so that redundant partitioning is avoided.

4.3.1 Controller input

The controller is designed to use any number of inputs, with any number of membership functions. During testing two environments were provided, one that uses the indoor and outdoor air temperatures as well as the relative humidity and CO_2 concentrations as inputs and another that does not use relative humidity.

These input variables were chosen since they are usually readily available (with a possible exception of the CO_2 concentration) and do provide the controller with sufficient information about its environment.

4.3.2 Controller output

The control variables are three. The first refers to the operating status of the heat pump and has seven possible settings, off and high, medium and low for heating and cooling respectively. The second is the air ventilation subsystem that can operate in three different modes, off, low and high. Finally the third is window control that can take one of the four following states: closed, slightly open, open, wide open. The resulting 84 possible actions are generated by the combinations of the above variables. The membership functions used for each of the output values are presented in the appendix.

The use of a heat pump model for both heating and cooling was preferred because of several advantages it offers without restricting the controller's applicability. Even if a building has separate heating and cooling systems they can be adequately simulated by the heat pump model since the

simultaneous operation of both systems would be inefficient and therefore avoided by the controller anyway. The use of one variable for both AC and heating leads to fewer possible actions (84 instead of 192 for the controller in question) that the RLFC needs to try and learn.

Further reduction of the possible actions can be achieved by restricting window usage during the operation of the AC system. This reduction could lead to as few as 30 actions at the expense of further complicating the controller design. Nevertheless, in the long term, the controller should learn what the best course of action is without the need of designer imposed restrictions.

The use of conventional heating bodies instead of a heat pump can be treated similarly although some modifications are essential. The control variable will have only five states: off, heating, cool low, cool medium and cool high. Although the reward assignment spreads over a number of time steps, the response delay of such a heating system can be quite large resulting in poor behavior. To overcome this problem we can add one more state referring to the current state of the heating bodies (cool, warm or hot). This added state will allow the controller to "forecast" the effect of turning on or off the heating.

4.4 The reinforcement learning linear controller (RLLC)

A second controller was also developed using linear function approximation of the state-action value function. The feature vector is constructed using radial basis functions (RBFs). RBFs were preferred to other ways of feature coding, because they provide continuous valued features using a simple and intuitive functional form. For example it is easy to determine the parameters of any number of RBFs that will evenly cover a given range. Besides that, there exist algorithms that can adjust the RBF parameters using supervised training. This feature can lead to better approximation but it was not used in this thesis because the resulting nonlinearities could lead to convergence problems. In some cases it has been found that RBFs with adaptive centers may leave parts of the space under-represented.

The RLLC approximates the state-action value function as the product of a weight vector and a feature vector. The exact construction of the feature vector will be discussed in a later section. The controller decides on the appropriate action by constructing the feature vector for every possible action under the current state. Then it multiplies all these vectors with the

weight vector and chooses the action with the largest expected value. It is therefore obvious that the controller performance depends on the accuracy of the weight vector. The weight vector is constructed at first randomly and consequently it is updated every time a new reward is received.

Using the TD(λ) algorithm we get the following update rule:

$$V_{t+1}(s_t) = V_t(s_t) + \alpha\delta e_{t+1}(s_t) \tag{4.5}$$

$$W_{t+1} = W_t + \alpha\left(r_t + \gamma\phi^T(x_{t+1})W_t - \phi^T(x_t)W_t\right)e_{t+1} \tag{4.6}$$

where $e_{t+1} = \gamma\lambda e_t + \phi(x)$.

The least squares problem as presented in equation (4.6) has the following objective function.

$$\begin{aligned} J &= \left\| \sum_{i=1}^{\infty} AW - \sum_{i=1}^{\infty} b \right\|^2 \\ A &= e_t\left(\phi^T(x_t) - \gamma\phi^T(x_{t+1})\right) \\ b &= e_t r_t \end{aligned} \tag{4.7}$$

Getting the least squares estimate of W requires a computation that involves the whole sequence of states, actions and rewards and has both computational and memory demands. In several applications of control adaptive filtering and system identification the recursive least squares algorithm (RLS) is used instead. The RLS algorithm updates the weight vector every time a training sample is available. The weight update rule as adapted for TD(λ) reinforcement learning by [Xu, et al., 23] is given by the following equations.

$$K_{t+1} = \frac{P_t e_t}{\left(\mu + \left(\phi^T(x_t) - \gamma\phi^T(x_{t+1})\right)P_t e_t\right)} \tag{4.8}$$

$$W_{t+1} = W_t + K_{t+1}\left(r_t - \left(\phi^T(x_t) - \gamma\phi^T(x_{t+1})\right)W_t\right) \tag{4.9}$$

$$P_{t+1} = \frac{1}{\mu}\left[P_t - P_t e_t\left[1 + \left(\phi^T(x_t) - \gamma\phi^T(x_{t+1})\right)P_t e_t\right]^{-1}\left(\phi^T(x_t) - \gamma\phi^T(x_{t+1})\right)P_t\right] \tag{4.10}$$

where $P_0 = \delta I$.

There are four tunable parameters in the RLS algorithm – δ, μ, γ and λ for the eligibility trace update. The δ parameter is used for the initialization of the P matrix and it has been shown that it can influence the convergence speed of the algorithm. The μ parameter is known from adaptive filtering as the forgetting factor and for the standard RLS-TD(λ) should be equal to

one. The γ and λ parameters are the same as those used in the TD(λ). It should be noted that no learning rate parameter is essential.

The RLLC action space is discrete and consists of the same actions available to RLFC.

4.5 Reinforcement signal design

In order for reinforcement learning to take place, a proper reinforcement signal is necessary. For the problem at hand the reinforcement signal should be a function of the energy consumption and the user satisfaction level. User satisfaction is further divided into thermal comfort and satisfaction with the indoor air quality.

The reinforcement signal is modeled as a variable that can take any value in the interval [-1 0]. This variable is in effect a penalty that is higher (closer to -1) during high energy consumption and/or user discomfort.

An estimate of current energy consumption can be obtained from the operational characteristics of the heating and cooling devices and their current operating settings. Estimating user satisfaction on the other hand is quite difficult using available measurements. Although we have a user response signal, this signal may not be used directly since it is irregular and therefore does not convey information regarding user satisfaction levels at every time step.

To overcome this problem an adaptive occupant satisfaction simulator (AOSS) was developed. The AOSS associates current environmental conditions inside and outside the building with user satisfaction level, so that for any given set of conditions AOSS' output will be 1 for user dissatisfaction and 0 for user satisfaction. Every time a signal from the user simulator is available, the AOSS is updated to incorporate the new information. In real building applications this information could be stored in an electronic card, so that when the user enters his or her office the environmental control will be suited to his or her preferences.

For the modeling of the indoor air quality a sigmoid of the CO_2 concentration is used that gives close to 0 values when the CO_2 concentration is less than 780 ppm and values close to 1 when the CO_2 rises above 950ppm.

The final reinforcement signal is given by the following equation:

$$\begin{aligned} \text{r.s.} = &-w_1 \cdot (\text{thermal comfort penalty}) - w_2 (\text{energy penalty}) \\ &-w_3 (\text{indoor air quality penalty}) \end{aligned} \tag{4.11}$$

$$\begin{aligned} \text{thermal comfort penalty} &= \frac{\sum_{t=0}^{k} \text{AOSS signal}}{k} \\ \text{energy penalty} &= \frac{\sum_{t=0}^{k} \text{energy consumption}}{\text{max energy consumption}} \\ \text{indoor air quality penalty} &= \frac{\sum_{t=0}^{k} \frac{1}{1+e^{[-0.06(CO_2 \text{ concentration} - 870)]}}}{k} \end{aligned} \tag{4.12}$$

The w_i variables are constants that represent the importance of each element, namely user satisfaction and energy conservation. Each constitute of the penalty is averaged over the time period between controller reevaluation so that the final reinforcement signal will be representative of the whole period.

In equation (4.12) the AOSS signal was used as measure of user dissatisfaction. The AOSS signal, as it has already been described, originates from the direct feedback of the building occupants. Of course using the AOSS is not always possible, especially when a large number of people share the same space. In such situations the use of the Fanger or the ACS model are preferable.

In order for the Fanger model to be used, the AOSS signal needs to be replaced with a value depicting discomfort, namely the PPD. Since the PPD takes values in the interval [0 1] and 0 corresponds to all user feeling comfortable and 1 all users being in discomfort, it is more suitable for direct use than the PMV index. Equivalently if the ACS model is in use the AOSS signal should be replaced by a measure of the distance of current indoor temperature to the comfortable one or by a binary feature showing if current indoor temperature is inside the comfort zone or not.

In a real building application, if the reinforcement signal is generated outside the controller, it is possible to change it. For example it is possible to use the PPD index and then replace it by an ACS index or a more accurate PPD. Since both controllers continually change their behavior

according to the received reward, it can be expected that the controllers will adapt to the new discomfort index.

From equation (4.12) it is obvious that the reinforcement signal will always be negative. Since we will set the initial rewards of the RLFC to zero the controller will search all possible actions at least once. This is because it will always find actions with higher penalty than it expects. If the controller is to implemented directly into a real building it would be advisable that the initial weights would reflect our expert knowledge so that the expected initial poor behavior is avoided.

4.6 Adaptive occupant satisfaction simulator (AOSS)

The AOSS was designed with the following things in mind. It should be a simple architecture, able to learn online, fast and without forgetting. Since simple feedforward neural networks are unsuitable for online training a reinforcement learning classifier was implemented.

This classifier uses reinforcement learning to classify current conditions as offending or non offending. The operation of the AOSS is the same to that of the RLFC, but for some minor differences. The action selection is greedy and the reward updating occurs according to the "short-sighted" algorithm. This means that eligibility traces are not used and that the values are updated towards the received reward and not towards a discounted return. The action selection was chosen to be greedy since we aim that eventually the AOSS will learn the user preferences and feedback from the user will no longer be required.

Since initially we have no information on the user preferences, the rewards for each action are chosen to be zero. The AOSS is trained using a generic user model. This model corresponds to a user that feels comfortable in the [-1.5 +1.5] PMV interval. This initial training is necessary to make the RLF-AOSS relatively accurate from the beginning so that the classifier will only need to learn the details of the real user's preferences.

4.7 Fis2con

A utility application was developed in Matlab C to aid controller design. This application creates the initial instance of the binary controller file that contains the fuzzy structure information and the initial rewards matrix. It uses as input a FIS structure that can be created using the user-friendly interface of the FIS Editor of the Fuzzy Toolbox for Matlab. The only

restriction posed is that the only membership functions it can process are triangular, trapezoid and gaussian.

The utility converts the membership functions described in the FIS structure to the three-parameter gaussian bells used by the controller. The three parameters are the mean, standard deviation and height of the bell. This is done to overcome the need of a controller that needs to parse different types of membership functions. Nevertheless the approximation error is in most cases small and it's influence on controller performance is negligible.

To convert the triangular membership functions, the bell of same mean and area is determined, while keeping the height unitary. The use of the height parameter is a trick so that the trapezoidal membership functions may be approximated. The intersecting sides of the trapezoid are extended until a triangle is formed. Using the aforementioned technique the triangle is approximated by a gaussian bell that has the same height as the triangle. Since the controller limits the output of all membership functions to the [0, 1] range, the associated approximation error will be small. The gaussian membership functions are left as they are with the height parameter unitary.

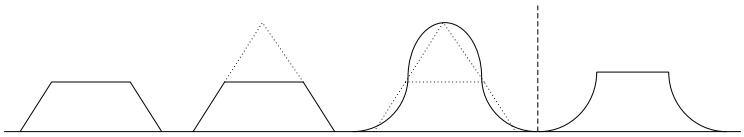


Figure 4-1: Conversion of trapezoidal membership functions to gaussian bells. First the trapezoid is converted to triangle and then to gaussian. The controller uses the limited gaussian on the right.

After all membership functions are processed the initial rewards matrix is appended to the file.

5 RESULTS AND CONCLUSIONS

5.1 Evaluation of the AOSS performance

In order to evaluate the AOSS performance a benchmark was developed in Simulink. This benchmark generates uniform random conditions in the entire range of environmental conditions. These conditions are then used to evaluate the PMV index according to ISO 7730 standard. Using the PMV index, user comfort is evaluated and a proper reinforcement signal is generated. Finally the controller given the random conditions and the reinforcement signal decides whether the user should be in comfort or discomfort and its decision is compared to the “real” one.

The input of the classifiers was chosen to be the relative humidity as well as the indoor and outdoor temperature. The user is considered to be in comfort when the PMV is in the [-1 +1] region.

The AOSS was tested several times using various degrees of input quantization. It was found that an increased amount of quantization provided a lower long-term error as expected. Of course after a certain limit the effect of the increased quantization was negligible and therefore could not justify the increased execution time. Figure 5-1 shows the error evolution for varying degrees of input quantization. All the classifiers were capable of reaching small errors in the long term even the one with only five membership functions per input.

The classifier was also tested using just two inputs, the indoor and outdoor temperature. The final error was between 4% and 8% when using 14 membership functions per input.

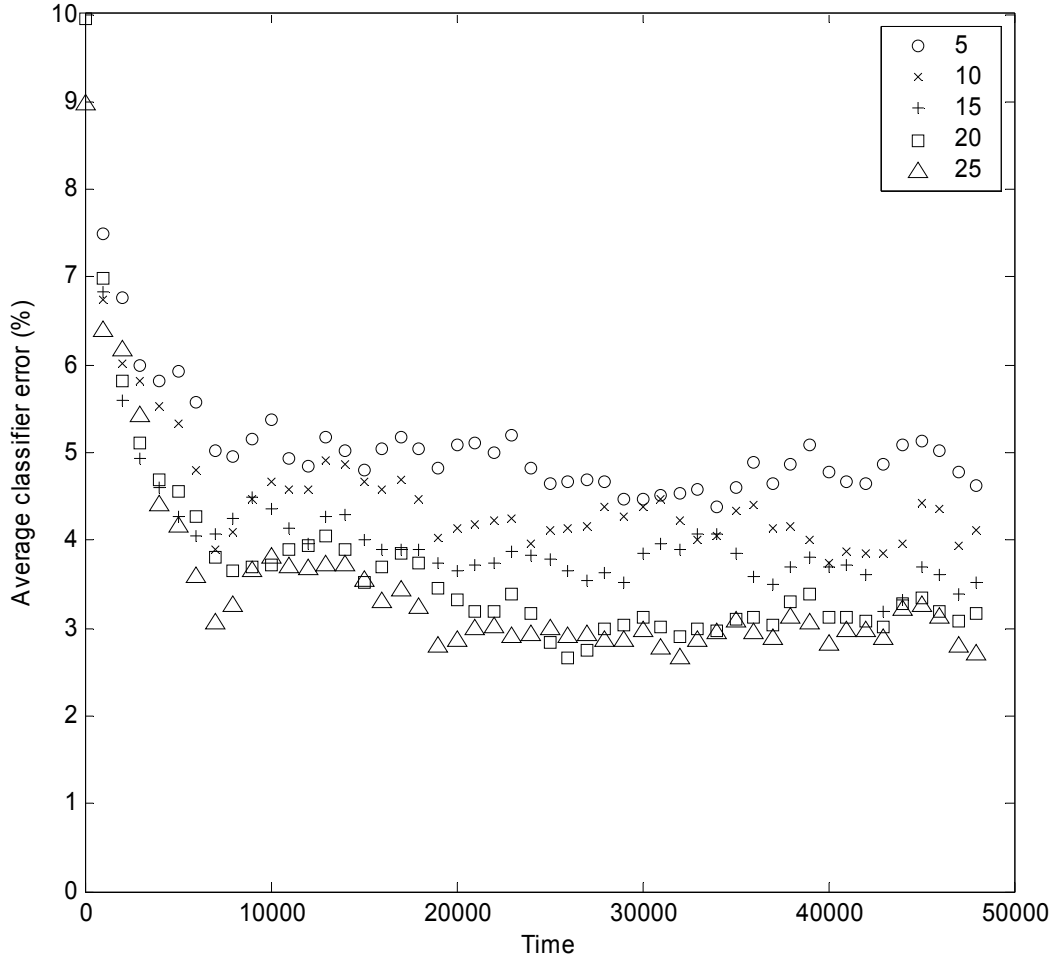


Figure 5-1: Classifier error as a function of time for various degrees of input quantization. The results are the average of five runs.

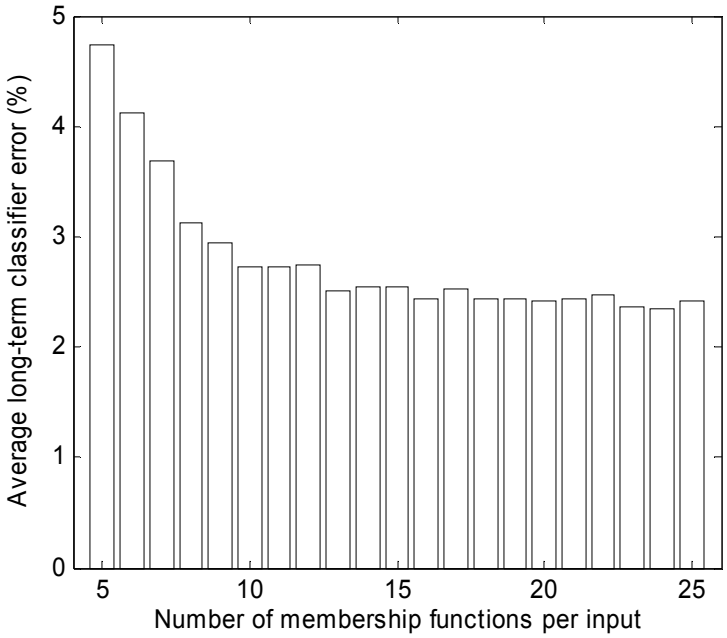


Figure 5-2: The dependence of long-term classifier error on the number of membership functions per input. The long-term error was calculated as the average classifier error during the last 20,000 of 50,000 steps. The results are the average of five runs.

A very important feature of the online classifier is its training speed. The training speed of the AOSS can be seen in Figure 5-3 where the first 3000 steps of an error curve are shown. Initially the error is close to 30% but just under 250 steps the error drops below 10%. 400 steps later the error begins to become steady around 5%. The learning curve did not vary significantly with increased quantization and in all cases the long-term error was reached in less than 1000 steps.

It should be noted that it took only 250 training samples to reach an error of 10% and less than 1000 to reach 5%. These samples unlike supervised offline training are introduced only once.

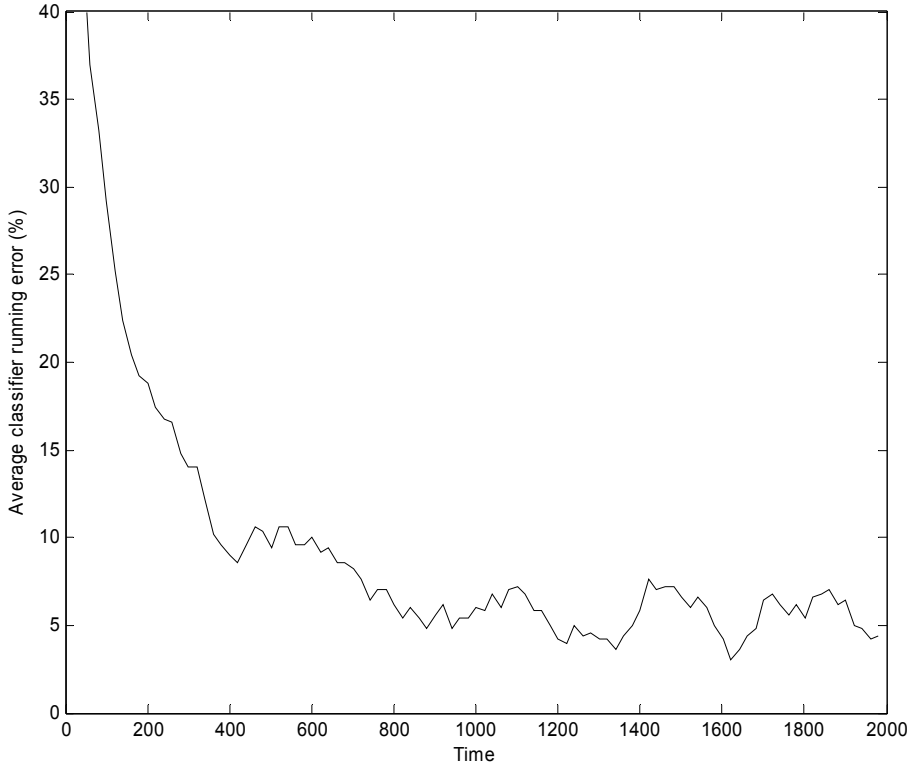


Figure 5-3: This figure depicts the classifier error for the first 2,000 steps for a RLF-AOSS with 10 membership functions per input. The results are averaged from five runs.

From Figure 5-4 we can determine the adverse effect of the number of membership functions per input to the speed of the AOSS. Specifically the average step execution speed remained unchanged for 5 to 8 membership functions but after that it increased dramatically. It is noteworthy that for 15 membership function execution time increased 208% and for 25 the increase was 440%.

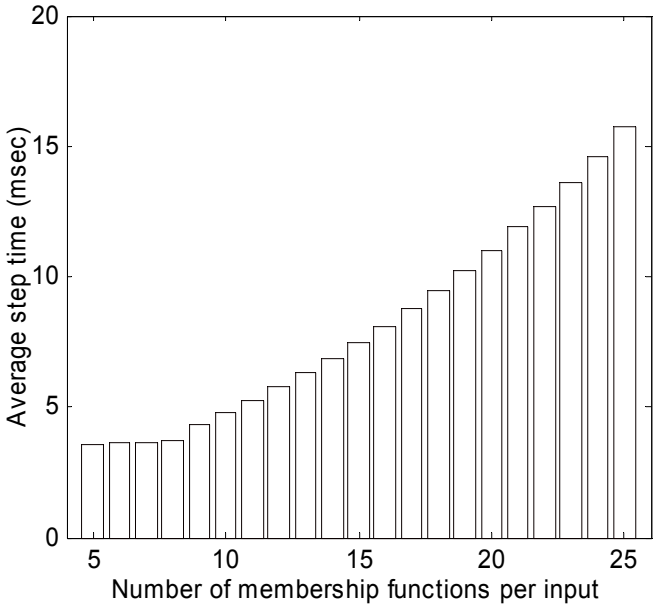


Figure 5-4: Execution time as a function of the number membership function per input. The times are averaged from 250,000 steps.

5.2 Controller Testing

In order to evaluate the performance of the controllers a suitable testing environment is required. This environment should incorporate a model of the building and one of the user responses. The building was simulated using the SIBIL application. SIBIL was developed by the Building Environmental Studies Group at the University of Athens and provides a parameterized Simulink model of a building.

Since user input is essential for the operation and evaluation of the controllers an add-in, called herein user simulator, was developed. This user simulator models user response based on current PMV conditions inside the building and the following tunable parameters:

Preference – Depending on climatic conditions and cultural background a user may prefer colder or warmer conditions, namely higher or lower PMV values.

Sensitivity – This parameter describes how far the PMV index can drift from the optimal value before the user senses discomfort.

Interest – This takes into account that each user will report his discomfort to the controller with different frequency.

There is also a facility to save user preferences along with current controller knowledge. If there is a need, the controller can be reset in order to start learning from scratch.

User response is modeled as a two state signal. One state denotes that the user feels discomfort, while the other provides no real information since it may denote that the current conditions are satisfactory or that although the opposite is valid the user did not report it.

Two buildings were modeled in the Sibil application. Building A has an area of $15m^2$, one window ($1m^2$) and the walls are made from 21cm of concrete and a 2,5cm insulating layer of foamed polystyrene. Building B has an area of $14m^2$ and one window ($2m^2$) facing in a different direction. The walls are made of 21cm of concrete but without insulation.

For comparison purposes all controllers are based on the PMV model and the average PPD index is used as a measure of thermal comfort.

5.3 Reference controllers

Two reference controllers were used. The first is an On/Off controller that uses the PMV index. This controller only operates the heating and cooling. Specifically it turns on the appropriate device when the PMV index moves outside the [-0.8 0.8] region and turns it off when the PMV moves inside the [-0.5 0.5] region. The ventilator unit is always on in the low setting to prevent large increase in CO_2 concentration levels. The response of this controller for a typical winter and summer 24 hour interval is shown in Figure 5-5. Since this controller operates based on the real PMV value it is expected that it will achieve low average PPD. In a real building application we should expect that the PMV index is only estimated or even that the controller will operate based on temperature. In any case it is doubtful that we can expect less energy consumption or smaller average PPD.

The second controller is a fuzzy-PD controller that is described in detail in [Kolokotsa, et al., 6]. This controller operates besides heating and cooling, the window and the ventilator. The response of this controller for a typical winter and summer day is shown in Figure 5-6.

The annual energy consumption of the On/Off controller in building A is about 4.77MWh and for building B is 8.65MWh. The corresponding consumptions for the fuzzy-PD controller are 3.28MWh and 5.83MWh respectively. During this year the On/Off controller achieved an annual average PPD of 13.4% and 16.7% for the second building while the fuzzy controller had 16.5% and 24.5% respectively. The maximum, minimum and mean CO_2 concentrations are summarized in Table 5-1.

Table 5-1: CO_2 concentration statistics for the On/Off and fuzzy-PD controllers.

	On/Off	**Fuzzy-PD**	
	Building A, B	**Building A**	**Building B**
Minimum	485	489	400
Mean	823	787	658
Maximum	1099	1098	935

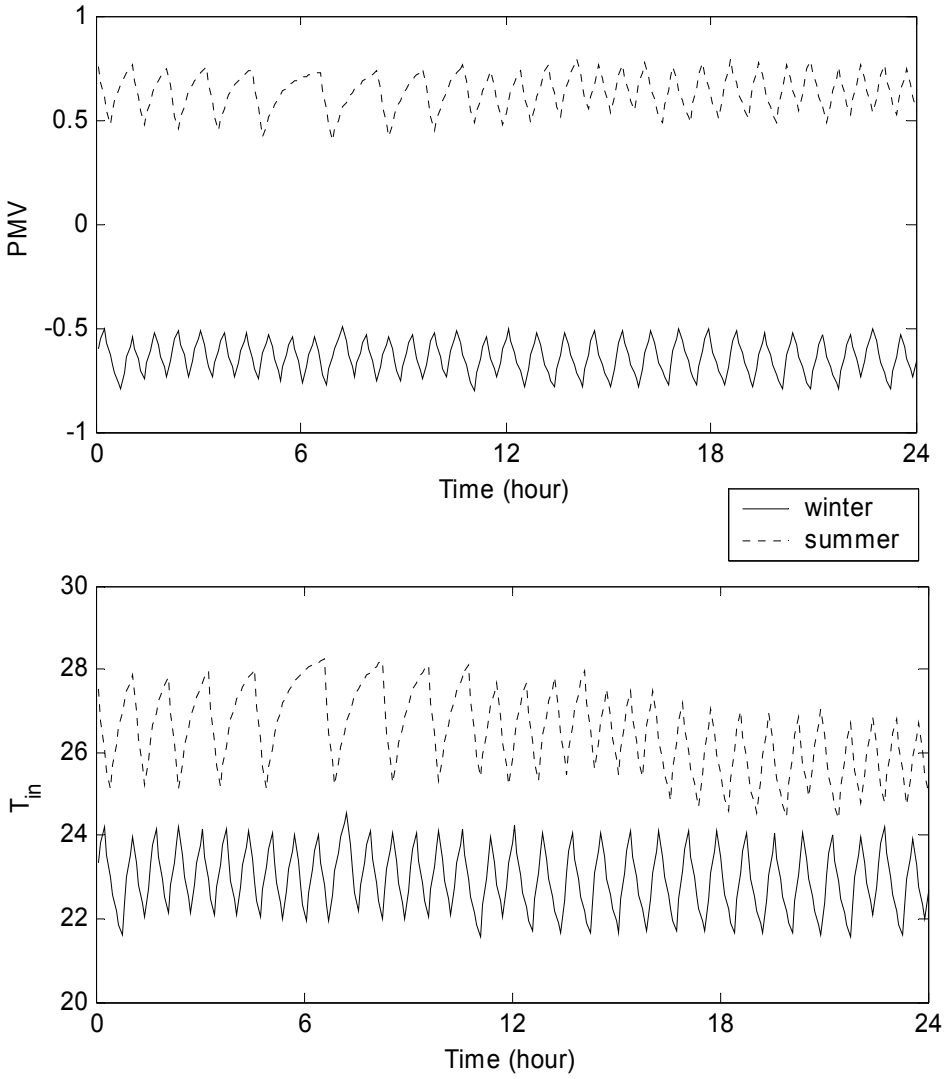


Figure 5-5: PMV and indoor temperature response of the On/Off controller for a typical winter and summer day.

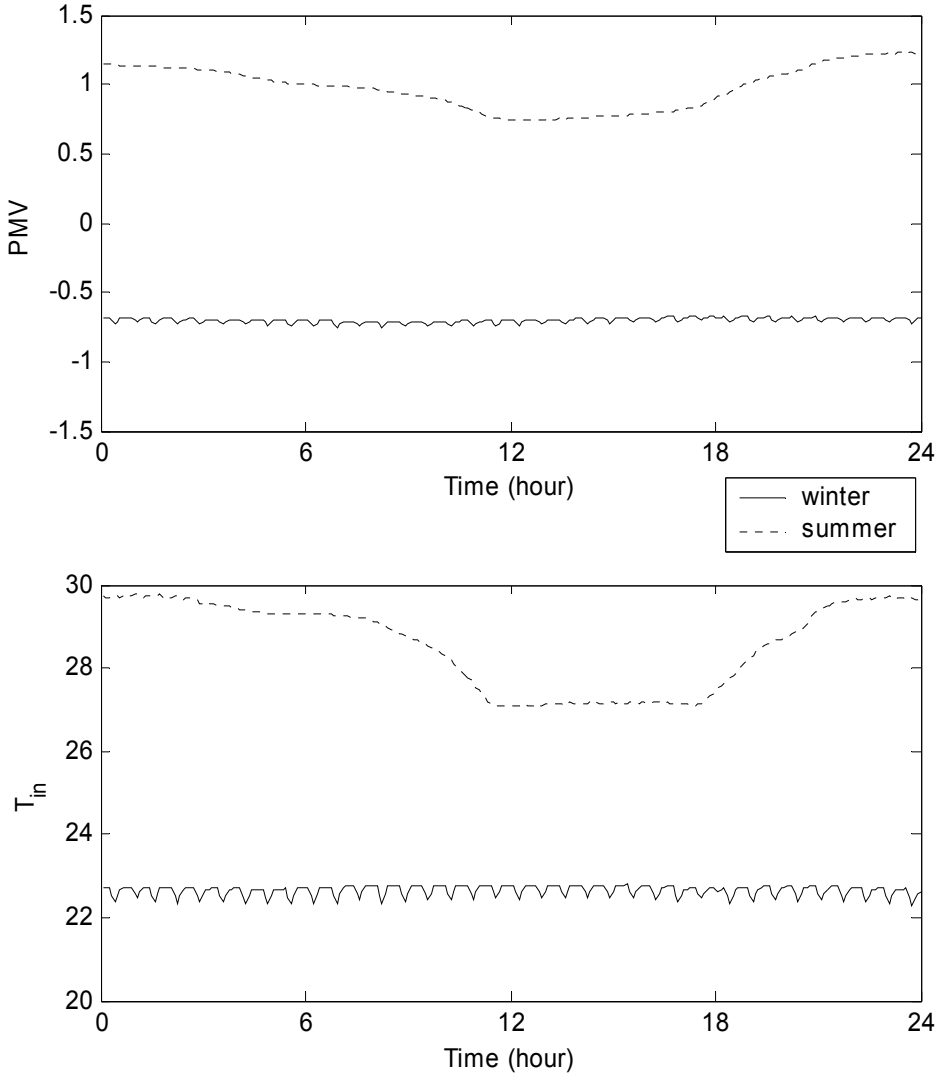


Figure 5-6: PMV and indoor temperature response of the fuzzy-PD controller for a typical winter and summer day.

5.4 RLFC Testing

5.4.1 Test configurations

The RLFC is tested using four different controller configurations:

A. Four inputs (T_{in}, T_{out}, RH, $[CO_2]$)

B. Four inputs (T_{in}, T_{out}, Time, $[CO_2]$).

C. Three inputs (T_{in}, T_{out}, $[CO_2]$).

The controller was tested for varying periods of time and parameter values (ε and learning rate).

5.4.2 RLFC performance

Although during repeated simulations the controller exhibited some improvement in its performance, none of the controllers tested ever reached a satisfactory policy up until this thesis was prepared. Specifically the controllers exhibit frequent policy changes due to value updating. In order to compensate to these frequent changes adaptive learning rates where attempted of the form $\alpha = \alpha_0 e^{-kt}$ but with no success.

5.5 RLLC Testing

5.5.1 Test configurations

The RLLC is tested using three different controller configurations:

A. Four inputs (T_{in}, T_{out}, RH, $[CO_2]$).

B. Four inputs (T_{in}, T_{out}, Time, $[CO_2]$).

C. Three inputs (T_{in}, T_{out}, $[CO_2]$).

For each of these configurations two different feature vectors were tested. The first feature vector consists of the values of all the state vectors, augmented by the action vector. The components of the action vector represent heating, cooling, ventilation and window opening as values in the [0, 1] range. The second feature vector comes from multiplying the first vector with the action vector. The smaller feature vector used was of size 34 and the largest of size 232.

The controller was tested for varying periods of time, RBFs and parameter values (ε, γ, λ). After testing the eligibility trace decay parameter was chosen to be 0.5. This value is consistent with what we expected, that is the actions taken up to 30 or 40 minutes ago should influence the current reward. The discount factors used are between 0.8 and 0.95 since higher

values are not recommended from the bibliography and smaller values (0.3 to 0.6) exhibited inefficient behavior. The inadequate behavior for low discount factors can be attributed to the fact that the controller probably found ways to increase immediate rewards, while simultaneously losing access to better long term rewards. It is possible that even by using small discount factors the controller will eventually converge to a good policy. The forget factor was chosen to be one at all cases since even a small change (0.99) caused bad behavior by the controller.

In general large feature vectors exhibited better performance as it was expected. Using feature vectors that combined states and actions resulted in increased performance since these feature vectors where able to capture some of the nonlinear relationships between states and/or actions.

5.5.2 RLLC performance

The RLLC controller exhibits adequate training speeds. It is noteworthy that even during the first year the controller is able to quickly develop a policy that although is far from optimal, it contains only few clearly wrong actions.

Applying an RLLC of the C configuration in building B we took the response depicted in Figure 5-7. This figure shows the controller's heat pump response with the exploratory actions eliminated. It is apparent that the controller quickly found that a good action during winter is to turn on heating and cooling during summer. It should be noted that the ε parameter was only 2% and that the exploratory actions where not completely random but chosen as one setting higher or lower than the calculated optimal. The annual energy consumption was 6.95MWh and the average PPD 31.5%. The high PPD is due to the fact that the controller begins with no knowledge of its environment and therefore makes a lot of mistakes especially in the beginning. This is evident from the fact that the average PPD of the last six months is only 25.5% while the average PPD of the first three months is more than 60%. The rest of the training parameters are summarized in Table 5-2.

Table 5-2: Training parameters of an one-year RLLC simulation

w_1	0.80	λ	0.5
w_2	0.05	γ	0.9
w_3	0.25	μ	1

The evolution of training is described in the following paragraph where the results from four single-year simulations are discussed. The controller tested corresponds to configuration B and was simulated using the building A definition. The parameters used during these simulations are cited in Table 5-3. The results are summarized in Table 5-4.

Table 5-3: Reinforcement learning parameters used for training a RLLC over a period of four years.

	1st year	2nd year	3rd year	4th year
w_1	0.80	0.80	0.80	0.80
w_2	0.01	0.01	0.01	0.01
w_3	0.20	0.20	0.20	0.27
λ	0.50	0.50	0.50	0.5
γ	0.95	0.95	0.95	0.90
μ	1.00	1.00	1.00	1.00
ε	0.075	0.025	0.000	0.000

Table 5-4: Results of simulating the RLLC for four years.

	1st year	2nd year	3rd year	4th year
Average PPD	12.1%	12.4%	12.8%	12.0%
Annual energy consumption	9.39MWh	7.13MWh	5.83MWh	4.85MWh
Minimum CO_2 concentration	385ppm	385ppm	389ppm	394ppm
Mean CO_2 concentration	464ppm	462ppm	450ppm	539ppm
Maximum CO_2 concentration	1658ppm	860ppm	860ppm	1697ppm

The results show that the annual average PPD does not change significantly with time but the energy consumption is reduced significantly from year to year by about 25% each time. At the same time the CO_2 concentrations vary within acceptable ranges after the first year, despite the fact that the CO_2 weight on the reinforcement signal is very small. It is noteworthy that even during the second year the CO_2 concentration is above 800ppm for less than an hour in a whole year. During the fourth and last year we increased the energy weight on the reinforcement signal and decreased the discounting factor. This had as an effect a decrease in the annual energy consumption and an increase of the CO_2 concentrations. The

latter can be attributed to the fact that the CO_2 concentrations are reversely analogous to the energy consumption. In order to conserve energy, the agent needs to reduce heat losses, closes the window and as a result the CO_2 concentrations increase.

Figure 5-8 shows the controller heat pump response for the first year and the corresponding PMV. Although a pattern is visible, there is a large number of random actions where the controller continuously switches from heating to cooling regardless of the season. This is due to the fact that the controller has no experience yet and because the ε value is 7.5% which means that the controller takes random actions quite frequently. Figure 5-9 shows the response of the controller during the second year of simulation. Now the ε value is smaller (2.5%) and the controller choices are based mostly on experience. Correspondingly the variations of the PMV index are smaller. Figure 5-10 corresponds to the third year simulated. This time the controller uses the greedy algorithm. It is obvious that it has learned not to use cooling during the winter months and although it occasionally chooses to turn heating during summer the performance is greatly improved.

The three figures described above provide only a very rough view of the controller's response. In order to better visualize the controller's true response, Figure 5-11 shows the variations of indoor and outdoor temperature and the PMV for a period of three days in winter and Figure 5-12 for a corresponding period during summer. The data used are from the third year of simulation. During the winter three day period the controller kept the indoor temperature at a mean value of 23.1ºC. The width of variation for the same period was 1.5ºC for the indoor temperature and 6.1ºC for the outdoor. The worst PMV value is -0.47 and the average is -0.31. Equivalently for the summer period the controller kept the temperature 26.5ºC with a variation width of 1.6ºC, while the outdoor temperature had a variation width of 9.1ºC. The worst PMV value is 0.91 with a mean of 0.73. The variations in indoor temperature during noontime occur due to the fact that the controller switches the cooling between low, medium and high in order to keep the temperature from rising while at the same time maintaining low energy consumption.

It should be noted that despite the fact that during the last year the greedy algorithm is used, the controller still learns and improves its performance by updating its value function.

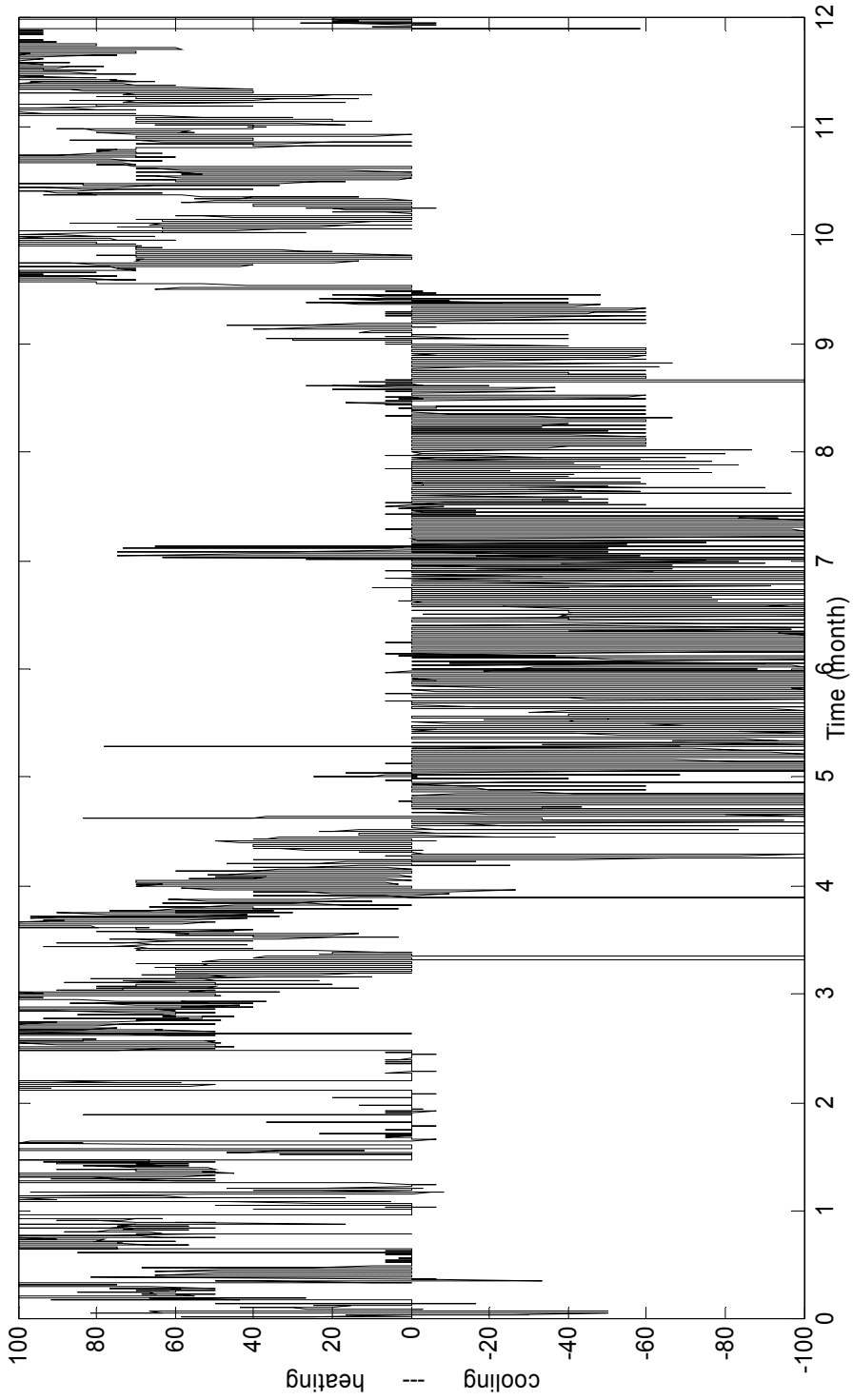


Figure 5-7: Heat pump response of the RLLC during its first year of simulated training. The response is averaged over an one hour period. This controller utilizes only three inputs (T_{in}, T_{out} and $[CO_2]$) and is applied in a building with no insulation.

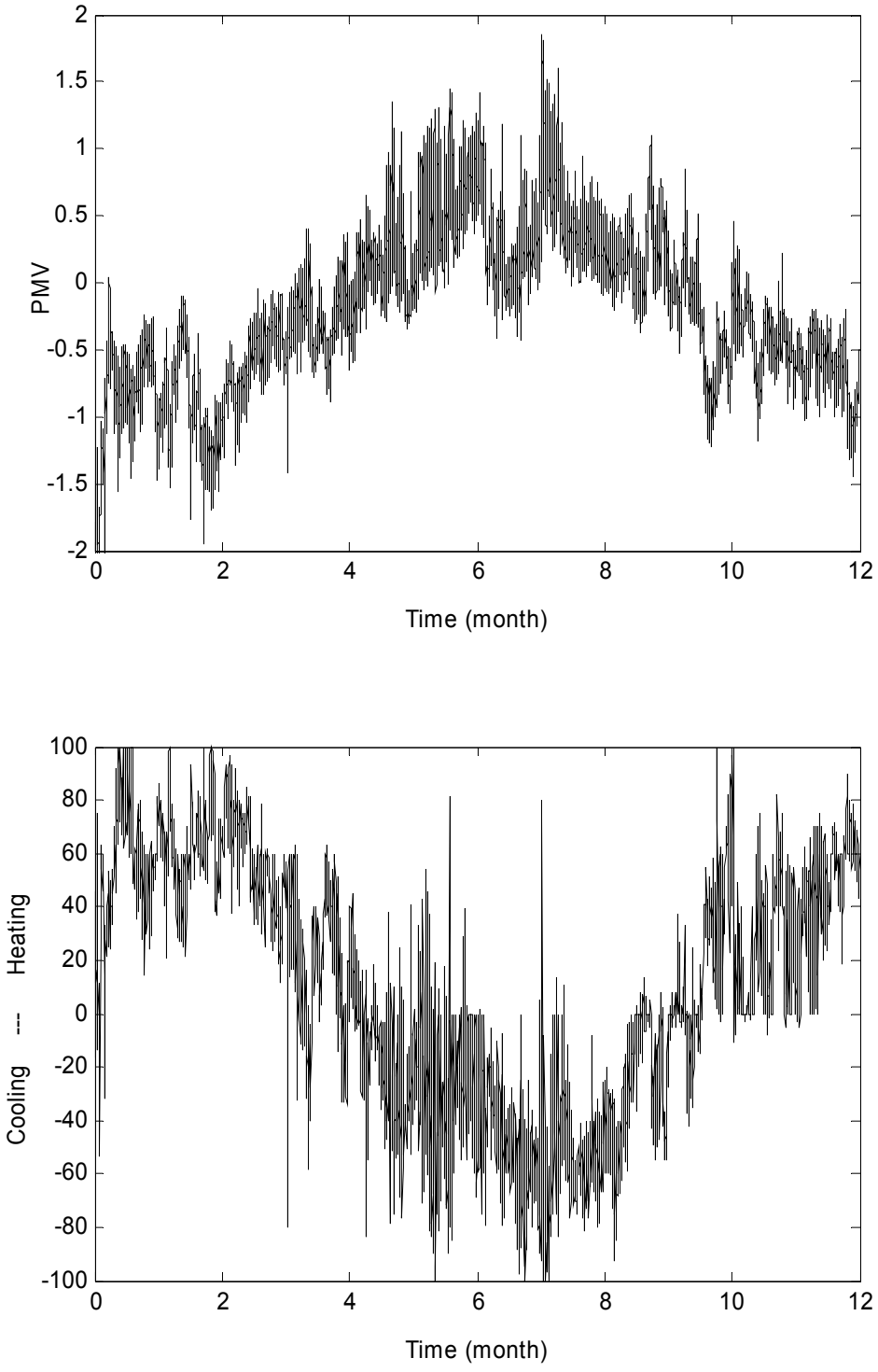


Figure 5-8: First year of RLLC simulated training. This controller utilizes four inputs (T_{in}, T_{out}, month and $[CO_2]$) and is applied in an insulated building. The heat pump response is averaged over a period of two hours.

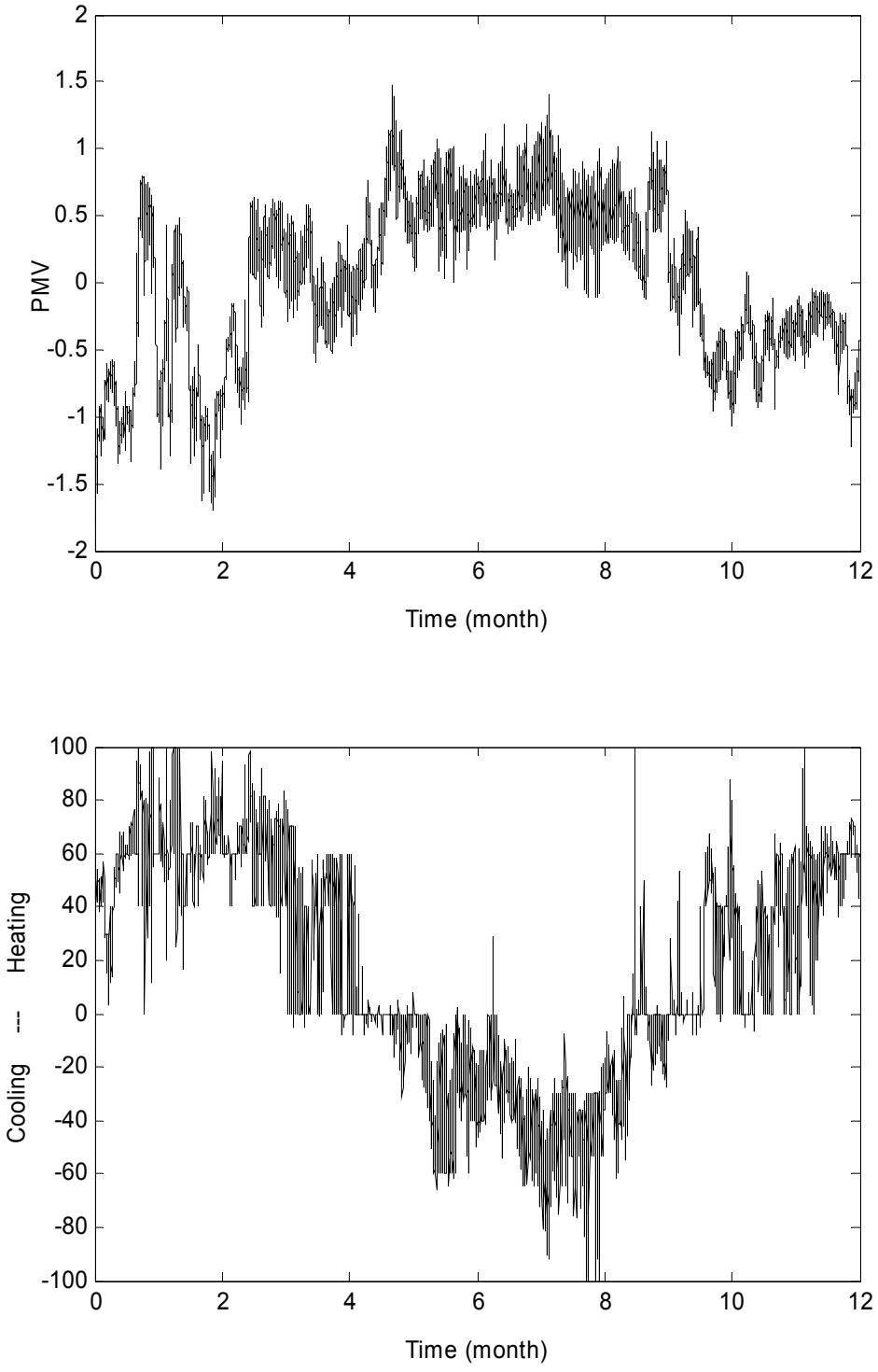


Figure 5-9: Second year of RLLC simulated training. This controller utilizes four inputs (T_{in}, T_{out}, month and $[CO_2]$) and is applied in an insulated building. The heat pump response is averaged over a period of two hours.

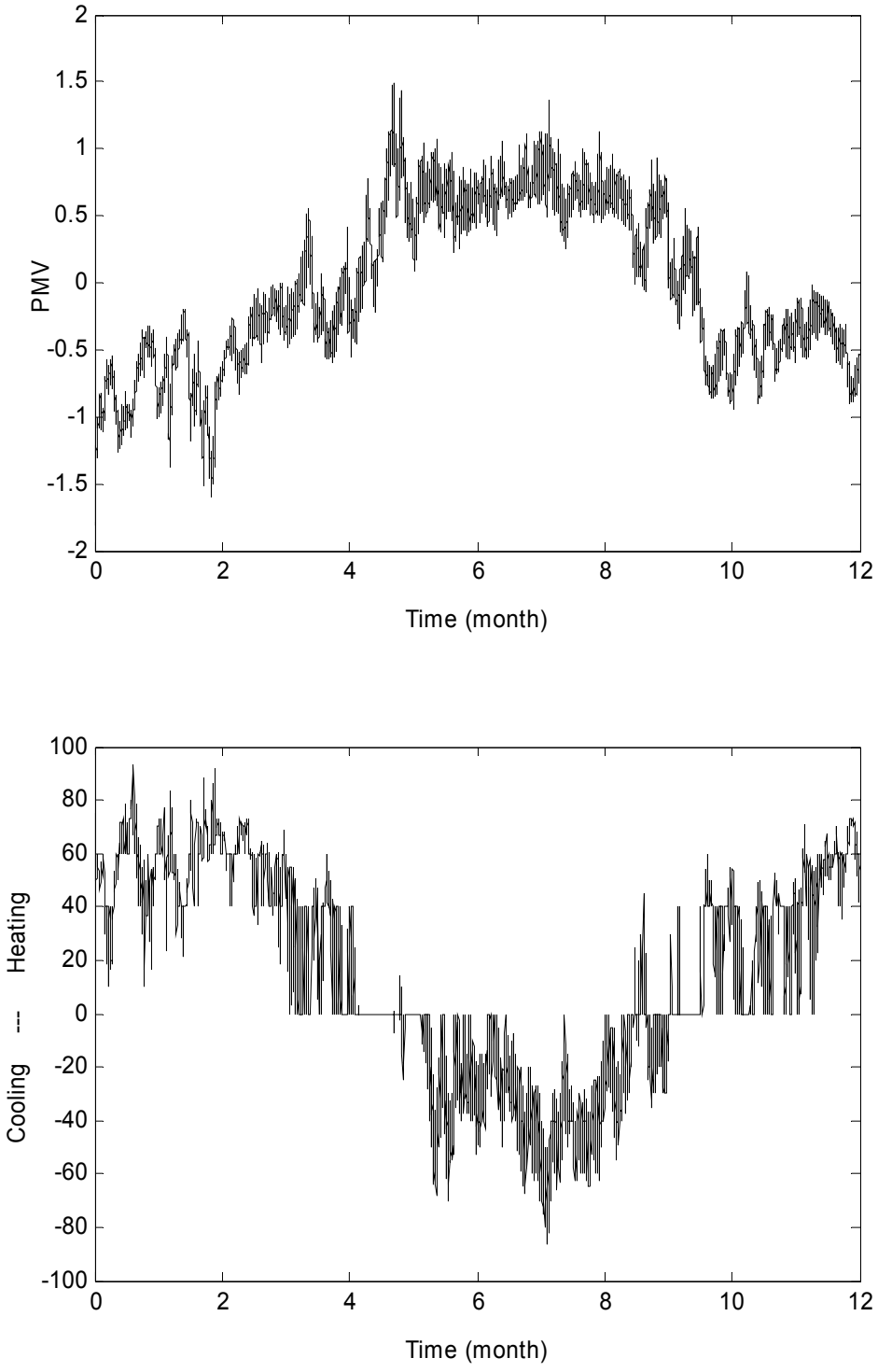


Figure 5-10: Third year of RLLC simulated training. This controller utilizes four inputs (T_{in}, T_{out}, month and $[CO_2]$) and is applied in an insulated building. The heat pump response is averaged over a period of two hours.

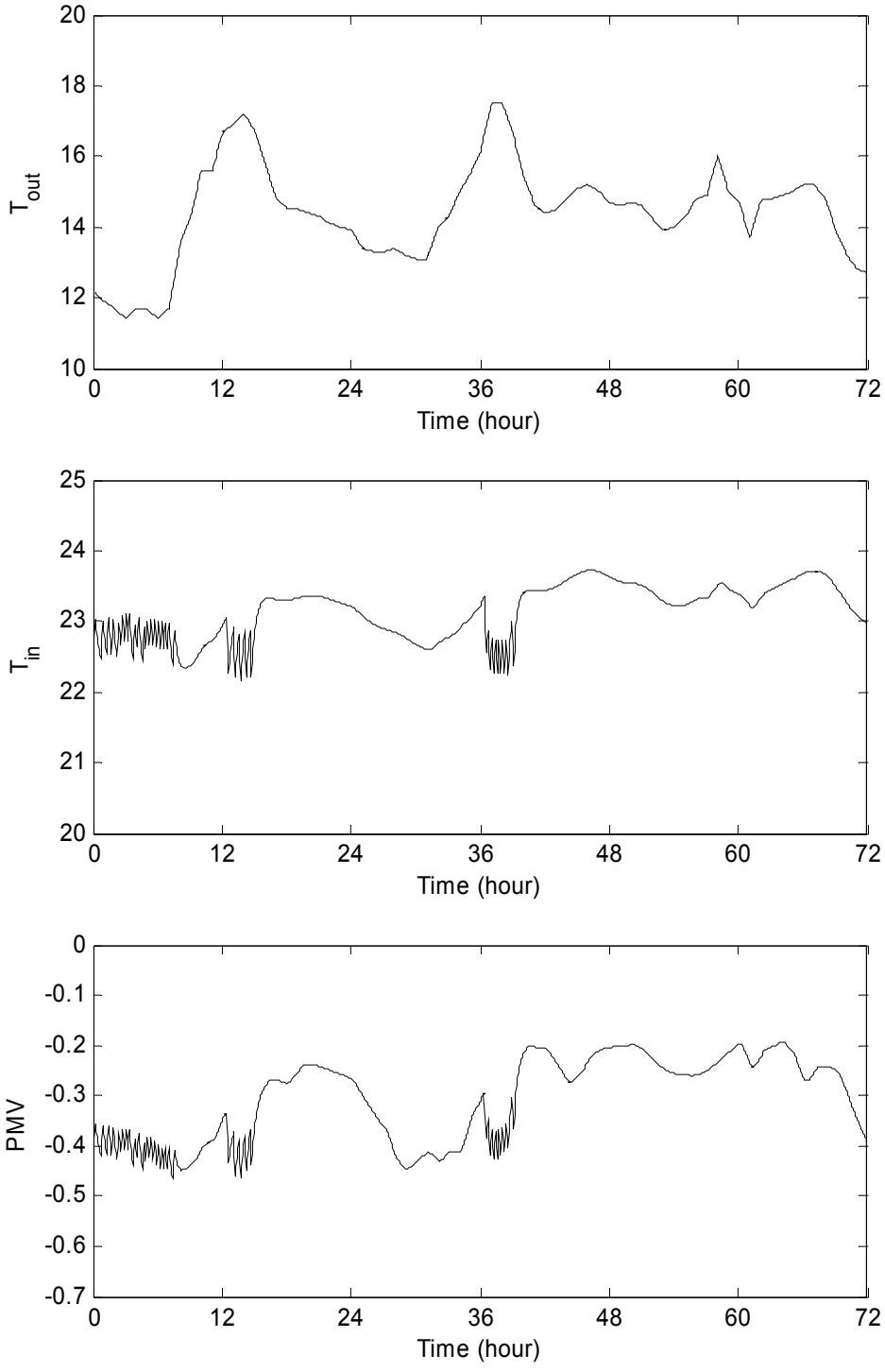


Figure 5-11: Temperature and PMV variations for a three-day winter period of a trained RLLC.

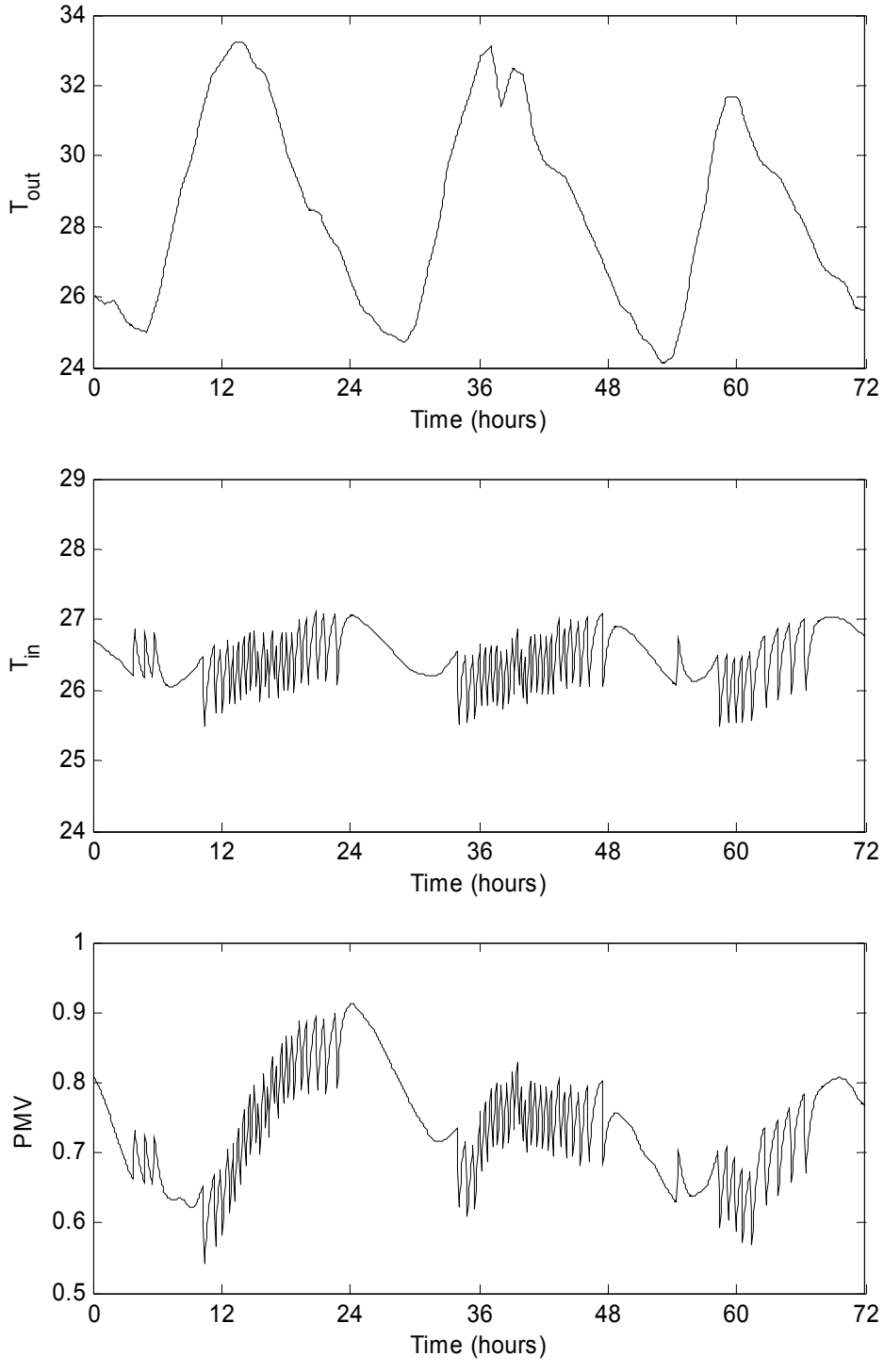


Figure 5-12: Temperature and PMV variations for a three-day summer period of a trained RLLC.

5.6 Conclusions

After training a RLLC for a simulated period of four years, we achieved the results summarized in Table 5-5 which also provides the corresponding results of the On/Off and Fuzzy-PD controllers. It is evident from this table that the RLLC has achieved a performance close to that of the other two controllers.

Table 5-5: Comparison of RLLC, On/Off and fuzzy-PD controllers

	RLLC	**On/Off**	**Fuzzy-PD**
Energy consumption	4.85MWh	4.77MWh	3.28MWh
Average PPD	12.0%	13.4%	16.5%

It is significant that the RLLC has achieved this performance while still making errors. Even during the fourth year the agent may turn heating in summer or cooling during winter. Further training should eliminate these wrong decisions and result in even better energy conservation. Significant improvement may also be achieved by using different feature vectors or by providing the controller with more information about its environment.

The main problem about the RLFC is that the initial choice of the state-action value matrix has a great influence on the convergence of the learning algorithm. This problem is more significant if we use greedy action selection or when non greedy actions are rare. The value matrixes used by the controllers tested had from 10,000 to 100,000 elements. A large percentage of these values are not actually used since they refer to states not actually encountered (40-90% depending on the inputs).

An expert choice of the initial value matrix is theoretically possible for the problem at hand but the large number of values makes it very difficult. Instead of choosing specific values for each state-action value, all values can be initialized as zero. A zero initial value has the effect of forcing the controller to try at least once each available action, because the resulting returns are always negative. On the other hand this can delay convergence because the controller may need to test each action several times before the estimate of its value can be accurate.

The learning rate parameter plays a significant role in the RLFC's training speed. When using the RLFC a proper learning rate must be determined usually as a function of time. If the learning rate does not become smaller with time it is possible that we will have frequent policy changes due to the value updating. On the other hand if the learning rate

drops too fast, the values may never reach the real ones and as a consequence the policy will be suboptimal.

The controllers are also compared according to their execution speed. It has been found that the RLFC is on average about 10% faster than the RLLC mostly because of the more complicated update step of the RLLC which requires matrix multiplication. The exact execution speed depends on the design characteristics of each controller. Of course for the application in study this time is negligible because the environment dynamics are very slow and the intervals between action reevaluation are large.

An issue that involves the application of reinforcement learning controllers in BEMS is that of sufficient exploration. It is true that taking random actions even during a small fraction of the time is unacceptable in a real building. Even when the choice is between near-optimal actions we should expect temporary increases in user dissatisfaction and an increase in the total energy consumption (2-3% for ε=0.02).

6 CONTRIBUTION – FUTURE ISSUES

6.1 Contribution

The main contribution of this thesis is that it has demonstrated the possibility to use reinforcement learning for building controller design. Reinforcement learning makes it possible to create controllers that can operate sufficiently with limited information, limited action possibilities and in changing environments. It is noteworthy that the controllers designed were able to function adequately using only indoor and outdoor temperature which are usually already available to the thermostatic controllers installed in a large number of dwellings and office spaces.

Besides that an innovative online classifier was developed that is also based on reinforcement learning. This classifier performed adequately in classifying the nonlinear relationship between environmental conditions and thermal comfort based only on relative humidity, indoor and outdoor temperature. The feature of online, fast training makes this classifier a candidate for several other applications.

6.2 Issues and future proposals

This thesis has shown that reinforcement learning controllers are feasible for BEMS. What has to be done now is to optimize these controllers so that they are able to adapt quickly and accurately. Convergence to an optimal policy is generally difficult to prove for reinforcement learning controllers. In the case of the controllers designed it is impossible since the state signal does not possess the Markov property. There follow some proposals for the improvement of the controllers.

6.2.1 Enhancing training speed

One of the most important problems encountered is that of the slow training speed. We have seen that the agent constantly improves its performance, but it is very important to achieve a near-optimal error-free policy as soon as possible.

A substantial increase in the controllers' training speed can be achieved by using state value functions instead of state-action value functions. Unfortunately in order to do that we need a way to determine the one step dynamics of the environment either deterministically or as transition probabilities. This means that we need a method to determine what the resulting state s´ will be from being in state s and choosing action a. The one-step dynamics can be obtained by a model of the environment, or by a second agent that estimates the dynamics of the environment based on past experience. Since a model of the environment cannot be assumed to be available the second solution is more appropriate.

To increase the controllers' training speed it is also possible to make a better usage of the available information. This can be achieved by updating the reward matrix or the weight vector more frequently. For example controller response can be evaluated every 15 or 20 minutes, which is usually sufficient, while the updating may occur every 5 or 3 minutes.

6.2.2 Using predictive control

Building control systems are known to benefit from the use of predictive control. Predicting the change in indoor environment in advance will provide additional information to assist action selection. Depending on the season and current environmental conditions, it is possible to predict the heat gains of a building and decide in advance how to act.

For example a decision of turning on heating can be the correct one during cold months and the opposite during hot months, although the conditions under which the action is taken can be the same. The

reinforcement learning controllers can learn to anticipate the changes in indoor environment by using additional input, for example the current month or the anticipated temperature from a weather forecast. Unfortunately the larger the dimensionality of the input, the slower the training speed of the controller.

6.2.3 Providing for artificial lighting

According to our design, the controller does not take into account the lighting requirements of the building and has no control over the artificial lighting and/or shading devices that might be available. This was chosen, so that the controller would not be burdened with additional inputs and action possibilities and the problems associated with the size of the resulting reward matrix or feature vector for RLFC and RLLC respectively.

In order to achieve the lighting requirements inside the building an independent controller is proposed. The controller used throughout the simulations was a simple fuzzy controller that is based on an earlier work by [Kolokotsa, et al., 6].

The use of shading may influence the thermal gains of the building especially in the summer period; this may result in poor behavior from the controller which would not be able to forecast the effect of its actions accurately. In order to alleviate this problem it is possible to select one of the controllers as the dominating or master controller and use its output to the other or slave controller.

A possible choice is to select the lighting controller as the master, so that shading will be controlled by it and the RLFC controller will have an additional input regarding the state of the shading device. For the initial testing of the controller it was decided that the influence of shading is minimal and therefore no connection between the two controllers was implemented.

6.2.4 Operation under faults

It has been discussed that reinforcement learning enables the controller to learn continuously even in changing conditions. This controller feature may result in problems in the case of faulty equipment. After the occurrence of a problem the controller will behave poorly and the conditions inside the building will probably drift from the optimal. If the fault persists then the controller will begin to adapt. For example if the window opening actuator breaks down the controller will begin to learn that

the window has no effect. This may result in unpredictable behavior during a period after the problem is fixed.

One way to prevent this from happening is the use of fault detection algorithms. If the algorithm is fast enough it can prevent the value updating of the controller during faults. Unfortunately this solution does not address the issue of poor controller behavior during the existence of the fault.

6.2.5 Night setback

Night setback refers to switching off the HVAC system during the time the building is unoccupied. This technique has been used frequently in the past to lower the energy consumption mainly in office buildings. Usually the system is turned back on some time before the occupants arrive, so that upon their arrival the indoor climate will be at the desired setpoints.

This effect can be implemented using the reinforcement learning controller. In order to do that the controller needs only to know the time. During the period that the building is empty the controller will not receive a thermal comfort or indoor air quality penalty. This way it will learn not to use the HVAC system during that period. Of course, upon occupant arrival the controller will receive a large penalty, if the conditions inside the building have drifted a lot away from optimum, therefore it will also learn to reestablish optimal conditions well in advance.

6.2.6 Giving control opportunities

As we discussed in chapter 2, when the building users have some control over their environment they tend to have wider comfort zones thus allowing for more energy conservation. It is possible to pass control of the operable windows from the controller to the occupants. In order to do that the controller only needs to know what the current status of the window is (e.g. open, half-open, closed). Using this information it will operate the rest of the available devices towards its goals.

7 APPENDIX

7.1 Metabolic rate (Met) values for various activities

Activity	Met
Reclining	0.8
Seated relaxed	1.0
Clock and watch repairer	1.1
Standing relaxed	1.2
Sedentary activity (office, dwelling, school, laboratory)	1.2
Car driving	1.4
Graphic profession - Book Binder	1.5
Standing, light activity (shopping, laboratory, light industry)	1.6
Teacher	1.6
Domestic work - shaving, washing and dressing	1.7
Walking on the level, 2 km/h	1.9
Standing, medium activity (shop assistant, domestic work)	2.0
Building industry -Brick laying (Block of 15.3 kg)	2.2
Washing dishes standing	2.5
Domestic work - raking leaves on the lawn	2.9
Domestic work - washing by hand and ironing (120-220 W/m2)	2.9
Iron and steel - ramming the mould with a pneumatic hammer	3.0
Building industry -forming the mould	3.1
Walking on the level, 5 km/h	3.4
Forestry - cutting across the grain with a one-man power saw	3.5
Agriculture - Ploughing with a team of horses	4.0
Building industry - loading a wheelbarrow with stones and mortar	4.7
Sports - Ice skating	6.2
Agriculture - digging with a spade (24 lifts/min.)	6.5
Sports - Skiing on level, good snow, 9 km/h	7.0
Forestry - working with an axe (weight 2 kg. 33 blows/min.)	8.6
Sports - Running, 15 km/h	9.5

7.2 Clo values for various garments

To estimate the total insulation effect of a person's clothing, sum the clo values of each garment. Also notice should be taken on the insulation effect of furniture like chairs or beds.

Garment description		Clo
Underwear, pants	Pantyhose	0.02
	Panties	0.03
	Briefs	0.04
	Pants 1/2 long legs, wool	0.06
	Pants long legs	0.1
Underwear, shirts	Bra	0.01
	Shirt sleeveless	0.06
	T-shirt	0.09
	Shirt with long sleeves	0.12
	Half-slip, nylon	0.14
Shirts	Tube top	0.06
	Short sleeve	0.09
	Light weight blouse, long sleeves	0.15
	Light weight, long sleeves	0.20
	Normal, long sleeves	0.25
	Flannel shirt, long sleeves	0.3
	Long sleeves, turtleneck blouse	0.34
Trousers	Shorts	0.06
	Walking shorts	0.11
	Light-weight trousers	0.20
	Normal trousers	0.25
	Flannel trousers	0.28
	Overalls	0.28
Coveralls	Daily wear, belted	0.49
	Work	0.50
Sweaters	Sleeveless vest	0.12
	Thin sweater	0.2
	Long sleeves, turtleneck (thin)	0.26
	Sweater 0.28 0.043 Thick sweater	0.35
	Long sleeves, turtleneck (thick)	0.37
Jacket	Vest	0.13
	Light summer jacket	0.25
	Jacket	0.35
	Smock	0.3

Garment description		Clo
Coats and overjackets and overtrousers	Coat	0.6
	Down jacket	0.55
	Parka	0.7
	Overalls multi-component	0.52
Highly-insulating coveralls	Multi-component, filling	1.03
	Fibre-pelt	1.13
Sundries	Socks	0.02
	Thick, ankle socks	0.05
	Thick, long socks	0.1
	Slippers, quilted fleece	0.03
	Shoes (thin soled)	0.02
	Shoes (thick soled)	0.04
	Boots 0.1 0.016 Gloves	0.05
Skirts, dresses	Light skirt, 15 cm. above knee	0.10
	Light skirt, 15 cm. below knee	0.18
	Heavy skirt, knee-length	0.25
	Light dress, sleeveless	0.25
	Winter dress, long sleeves	0.4
Sleepwear	Long sleeve, long gown	0.3
	Thin strap, short gown	0.15
	Hospital gown	0.31
	Long sleeve, long pyjamas	0.50
	Body sleep with feet	0.72
	Undershorts	0.1
Robes	Long sleeve, wrap, long	0.53
	Long sleeve, wrap, short	0.41
Chairs	Wooden or metal	0.00
	Fabric-covered, cushioned, swivel	0.10
	Armchair	0.20

7.3 Illuminance requirements

The following table shows general illuminance requirements depending on the activity. Modifying factors for the general requirements are also shown for specific situations.

Activity	Required illuminance
High eye strain activities: precision drawing, jewelry etc.	1000 lux
Short duration activities with high eye strain: reading, drawing etc.	750 lux
Short duration activities with medium eye strain: work in general, meetings etc.	500 lux
Short duration activities with low eye strain: storage, movement etc.	250 lux
×0.8 Age < 35years / Activity unimportant / Low difficulty	
×1.2 Age > 55 years / Activity crucial or unusual / High difficulty	

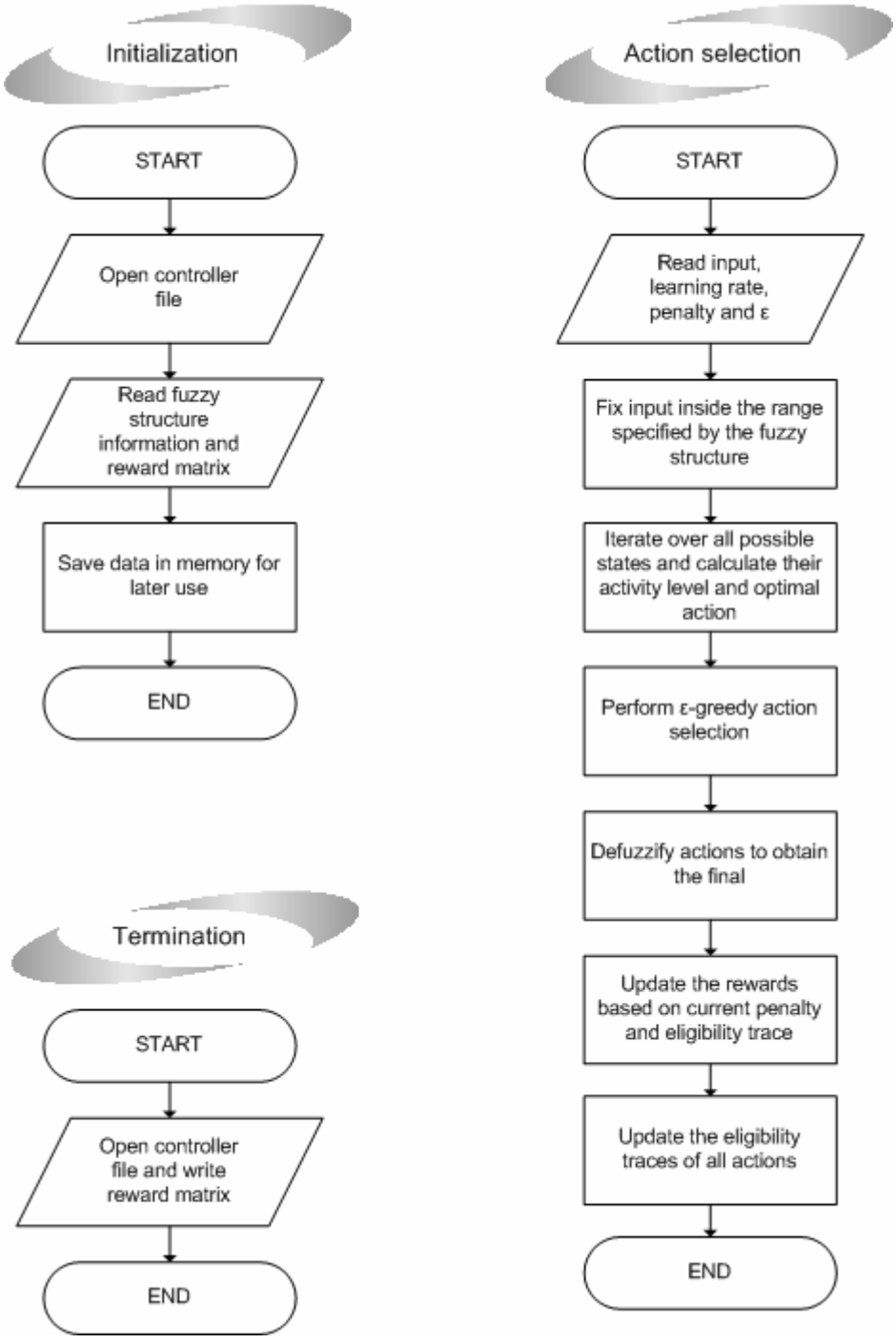
7.4 Glare requirements

Condition	Glare index
Highly critical conditions with difficult work, dangerous situations etc.	<13
Conditions with long-duration work of normal difficulty, with rest periods etc.	13-16
Conditions with short-duration work or light work, with long breaks etc.	16-19
Conditions below critical, with short work periods, movement etc.	19-22
Conditions without visual requirements, in which glare is not a problem	>22

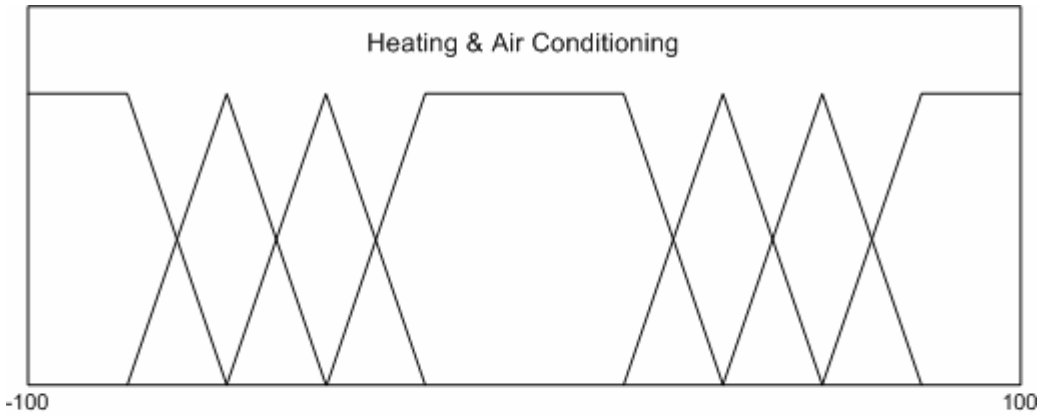
7.5 Light color

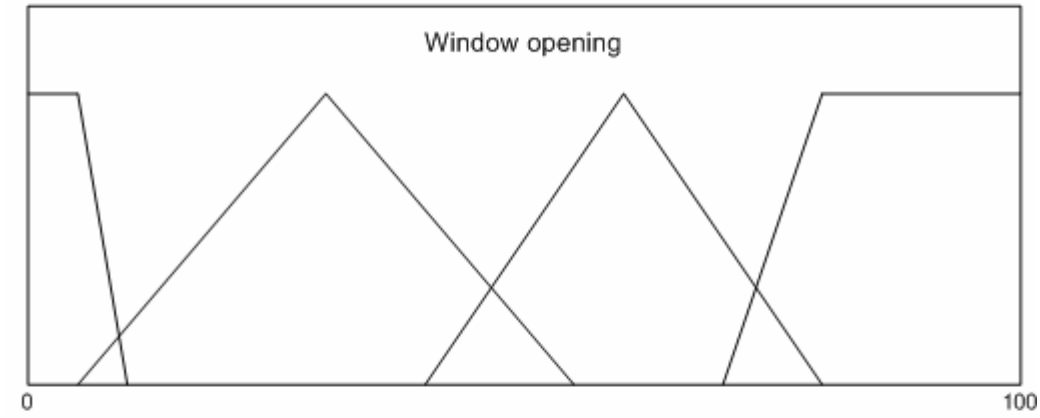
Type of space	Condition	T (K)
Spaces where color is very important	Work	4500-6000
	Rest	2500-4000
Spaces where color is important but not critical	Work	> 4000
	Rest	< 4000
Spaces where chromatic recognition is unimportant	Work	> 4500
	Rest	< 4500
Space without chromatic vision		Indifferent

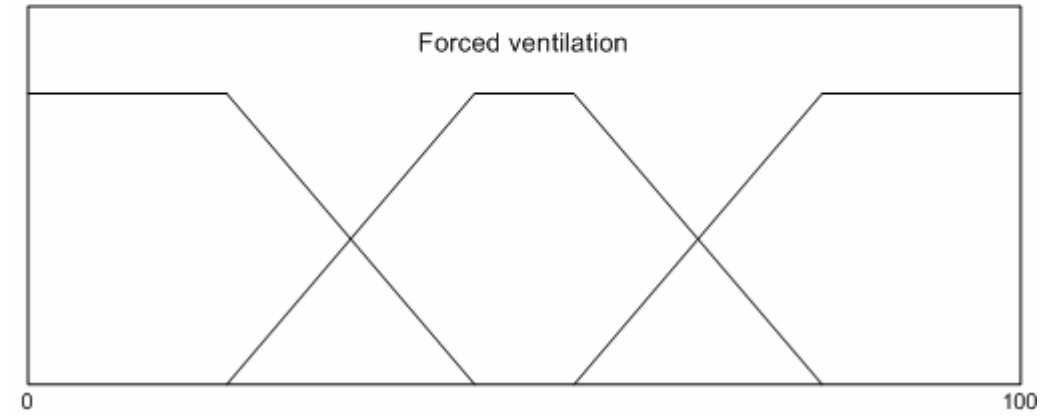
7.6 Reinforcement Learning Fuzzy Controller flowchart



7.7 Action membership functions







Each action is chosen in the range of 0-100 thus referring to the percentage of the intensity of each action. In the heating and air conditioning membership functions positive numbers refer to heating and negative numbers refer to cooling.

8 INDEX OF TABLES AND FIGURES

9 INDEX OF ABBREVIATIONS

ACS	Adaptive comfort standard
AMV	Actual mean vote
AOSS	Adaptive Occupant Satisfaction Simulator
ASHRAE	American Society of Heating, Refrigerating and Air-Conditioning Engineers
BEMS	Building Energy Management Systems
COA	Center of area defuzzification method
DP	Dynamic Programming
HVAC	Heating, Ventilation, Air-Conditioning
IAQ	Indoor Air Quality
ISO	International Standards Organization
MC	Monte Carlo
MDP	Markov Decision Process
MOM	Mean of max defuzzification method
MSE	Mean Squared Error
NC	Noise criteria
PMV	Predicted mean vote
PPD	Percent of people dissatisfied
RBF	Radial Basis Functions
RLFC	Reinforcement Learning Fuzzy Controller
RLLC	Reinforcement Learning Linear Controller
RLS	Recursive Least Squares
SBS	Sick Building Syndrome
TD	Temporal Difference
VOC	Volatile Organic Compounds

10 NOMENCLATURE

a	Action
e	Eligibility trace
E	Illuminance
I	Luminous intensity
J	Cost function
L	Luminance
m(s)	Membership – Activity level of state s
M	Metabolic rate
Q	State-Action value
R	Return
s	State
T_C	Comfort temperature
T_{in}	Indoor air temperature
T_{out}	Outdoor air temperature
T_{RM}	Running mean temperature
V	State value
w_1	Weight of thermal comfort penalty in reinforcement signal
w_2	Weight of energy penalty in reinforcement signal
w_3	Weight of indoor air quality penalty in reinforcement signal
W	Weight vector
α	Learning rate
γ	Discount parameter
δ	TD error
θ	Parameter vector
λ	Eligibility trace decay parameter
μ	Forgetting factor
σ	Standard deviation
Φ	Luminous flux
ϕ	Feature vector

11 REFERENCES

[1] "Building Energy Management Systems," in *Design and Maintainance Guide 22*: UK Ministry of Defence, Defence Estates, 2001, pp. 1-7.

[2] J. A. Clarke, J. Cockroft, S. Conner, J. W. Hand, N. J. Kelly, R. Moore, T. O'Brien, and P. Strachan, "Simulation-assisted control in building energy management systems," *Energy and Buildings*, vol. 34, pp. 933-940, 2002.

[3] M. Hamdi and G. Lachiver, "A fuzzy control system based on the human sensation of thermal comfort," pp. 487-492, 1998.

[4] P. Salgado, J. B. Cunha, and C. Couto, "A computer-based fuzzy temperature controller for environmental chambers," pp. 1151-1156.

[5] A. I. Dounis and D. E. Manolakis, "Design of a fuzzy system for living space thermal-comfort regulation," *Applied Energy*, vol. 69, pp. 119-144, 2001.

[6] D. Kolokotsa, D. Tsiavos, G. S. Stavrakakis, K. Kalaitzakis, and E. Antonidakis, "Advanced fuzzy logic controllers design and evaluation for buildings' occupants thermal - visual comfort and indoor air quality satisfaction," *Energy and Buildings*, vol. 33, pp. 531-543, 2001.

[7] A. E. Ben-Nakhi and M. A. Mahmoud, "Energy conservation in buildings through efficient A/C control using neural networks," *Applied Energy*, vol. 73, pp. 5-23, 2002.

[8] N. Morel, M. Bauer, M. El-Khoury, and J. Krauss, "Neurobat, a predictive and adaptive heating control system using artificial neural networks," *International Journal of Solar Energy*, vol. 21, pp. 161-201, 2001.

[9] B. Egilegor, J. P. Uribe, G. Arregi, E. Pradilla, and L. Susperregi, "A fuzzy control adapted by a neural network to maintain a dwelling within thermal comfort," *Proceedings of Building Simulation 97*, vol. 2, pp. 87-94, 1997.

[10] F. Yamada, K. Yonezawa, S. Sugarawa, and N. Nishimura, "Development of air-conditioning control algorithm for building energy-saving," presented at IEEE International Conference on Control Applications, Hawai, USA, 1999.

[11] Thermal Comfort, by Innova AirTech Instruments, www-impind.de.unifi.it/impind/didattica/materiale/microclima/innova/thermal.htm,1997, Accessed on 11/3/2003

[12] B. W. Olesen and K. C. Parsons, "Introduction to thermal comfort standards and to the proposed new version of EN ISO 7730," *Energy and Buildings*, vol. 34, pp. 537-548, 2002.

[13] F. Memarzadeh and A. Manning, "Thermal Comfort, Uniformity, and Ventilation Effectiveness in Patient Rooms: Performance Assessment Using Ventilation Indices," *ASHRAE Transactions: Symposia*, 2000.

[14] B. W. Jones, "Capabilities and limitations of thermal models for use in thermal comfort standards," *Energy and Buildings*, vol. 34, pp. 653-659, 2002.

[15] R. J. d. Dear and G. S. Brager, "Thermal comfort in naturally ventilated buildings: revisions to ASHRAE Standard 55," *Energy and Buildings*, vol. 34, pp. 549-561, 2002.

[16] M. A. Humphreys and J. F. Nicol, "The validity of ISO-PMV for predicting comfort votes in every-day thermal environments," *Energy and Buildings*, vol. 34, pp. 667-684, 2002.

[17] J. F. Nicol and M. A. Humphreys, "Adaptive thermal comfort and sustainable thermal standards for buildings," *Energy and Buildings*, vol. 34, pp. 563-572, 2002.

[18] K. J. McCartney and J. F. Nicol, "Developing an adaptive control algorithm for Europe," *Energy and Buildings*, vol. 34, pp. 623-635, 2002.

[19] R. Serra, "Chapter 6 - Daylighting," *Renewable and Sustainable Energy Reviews*, vol. 2, pp. 115-155, 1998.

[20] Carrier, "Total environmental quality," in *Synopsis*, vol. 3.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*: MIT Press, 1998.

[22] S. I. Reynolds, "Reinforcement learning with exploitation," in *School of Computer Science*. Birmingham: The University of Birmingham, 2002.

[23] X. Xu, H.-g. He, and D. Hu, "Efficient Reinforcement Learning Using Recursive Least-Squares Methods," *Journal of Artificial Research*, pp. 259-292, 2002.

Εφαρμογή ενισχυμένης μάθησης για άνεση και εξοικονόμηση ενέργειας σε κτίρια.

Η εργασία αυτή αναφέρεται στο πρόβλημα της επίτευξης άνεσης σε κτίρια με την ελάχιστη δυνατή κατανάλωση σε ενέργεια. Αν και θέμα έρευνας εδώ και δεκαετίες η θερμική άνεση παραμένει ανοιχτή υπόθεση. Τα διεθνή πρότυπα αναπροσαρμόζονται έτσι ώστε να συμπεριλάβουν το νέο πρότυπο της προσαρμοστικής άνεσης, ενώ ο μέχρι πρότινος καθιερωμένος δείκτης προβλεπόμενης μέσης ψήφου τίθεται υπό αμφισβήτηση όσον αφορά τη δυνατότητα εφαρμογής του σε κτίρια φυσικής ροής. Παράλληλα, η προσπάθεια για εξοικονόμηση ενέργειας που ξεκίνησε την δεκαετία του 1970, σε συνδυασμό με την απαίτηση για μείωση των εκπομπών διοξειδίου του άνθρακα, καθιστούν επιτακτική τη χρήση συστημάτων ενεργειακού ελέγχου των κτιρίων. Τις τελευταίες δεκαετίες οι εφαρμογές αυτομάτου ελέγχου έχουν ωφεληθεί πολύ από την χρήση της τεχνητής νοημοσύνης – νευρωνικά δίκτυα, ασαφής έλεγχος, ενισχυμένη μάθηση. Από τις πιο πρόσφατες, η τεχνική της ενισχυμένης μάθησης παρουσιάζει ιδιαίτερο ενδιαφέρον. Η ενισχυμένη μάθηση αναφέρεται στη μάθηση μέσω της αλληλεπίδρασης με το περιβάλλον. Ο ελεγκτής δεν γνωρίζει εκ των προτέρων ποιες είναι οι σωστές απαντήσεις. Σε αντίθεση με άλλες τεχνικές προσαρμοστικού ελέγχου, μετά από κάθε απόφαση που παίρνει δεν του δίνεται ποια ήταν η σωστή απάντηση, αλλά μόνο μία ένδειξη για το πόσο καλή ή κακή ήταν. Έτσι ο ελεγκτής προσπαθεί να παίρνει τέτοιες αποφάσεις ώστε να μεγιστοποιεί τις καλές αποκρίσεις. Στην εργασία αυτή αναπτύσσονται δύο ελεγκτές που βασίζονται στην ενισχυμένη μάθηση. Ο πρώτος κάνει χρήση ασαφούς λογικής για να διακριτοποιήσει την είσοδό του, ενώ η λειτουργία του δεύτερου βασίζεται σε γραμμική προσαρμογή με την επαναληπτική μέθοδο των ελαχίστων τετραγώνων και χρήση χαρακτηριστικού διανύσματος που δημιουργείται με radial basis functions. Το σήμα μάθησης που παίρνουν οι ελεγκτές είναι μια συνάρτηση της θερμικής άνεσης των χρηστών του κτιρίου, της ποιότητας της ατμόσφαιρας εντός του κτιρίου και της ενεργειακής κατανάλωσης. Και οι δύο ελεγκτές αναπτύχθηκαν και προσομοιώθηκαν στο περιβάλλον Matlab/Simulink, έτσι ώστε να υπολογιστεί η απόδοσή τους.

Reinforcement learning for energy conservation and comfort in buildings.

This thesis deals with the issue of achieving comfort in buildings with minimal energy consumption. Although the issue of comfort has been investigated for decades, thermal comfort remains an open issue. The international standards are reevaluated to include the new adaptive comfort standard, while the applicability of the PMV index in naturally ventilated buildings is under scrutiny. At the same time the effort for energy conservation that begun in the 1970s, along with the CO_2 emission reduction requirements, render the use of energy management systems in buildings imperative. During the last decades the applications of automatic control have profited from the use of artificial intelligence – neural networks, fuzzy logic and reinforcement learning. The technique of reinforcement learning is of particular interest. In contrast to other techniques of adaptive control, a reinforcement learning agent does not know what the correct answer is, instead it receives only an indication of the “correctness” of its response. For this thesis two reinforcement learning controllers have been developed. The first one makes use of fuzzy logic to quantize its input space, while the other one applies linear approximation using the recursive least squares method. The feature vector required for the latter is created using radial basis functions. The learning signal used for both controllers is a function of the thermal comfort of the building occupants, the indoor air quality and the energy consumption. Both controllers were developed and simulated in the Matlab/Simulink environment in order to assess their performance.